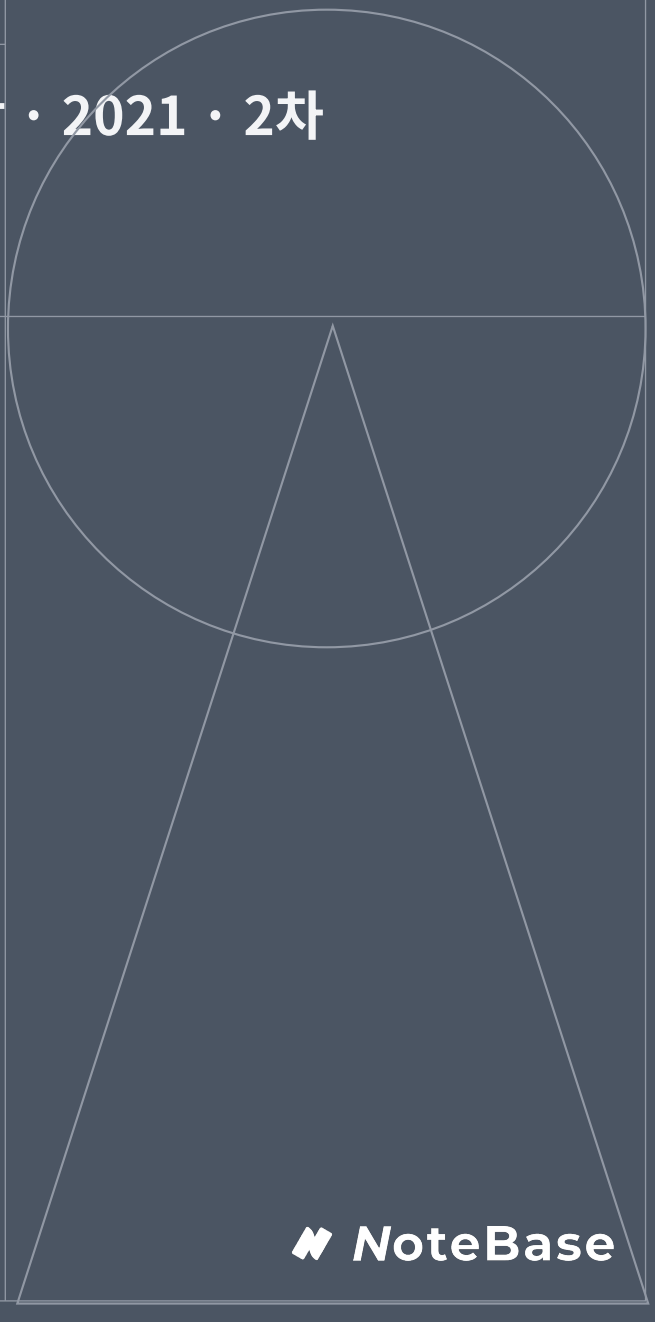
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2021 · 2차



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2021 · 2차

공무원/국가직/7급 · 통계학 · 2021 · 2차

문 1. 다음 사례에 해당하는 표본추출방법은?

서울특별시의 25개 각 구로부터 20개씩의 가구를 임의로 뽑아 가구주에 대해 환경문제에 관한 면접조사를 실시하기로 했다.

① 단순임의추출법(simple random sampling)
② 층화추출법(stratified sampling)
③ 계통추출법(systematic sampling)
④ 집락(군집)추출법(cluster sampling)

정답 2번

출제 포인트

제시된 사례가 단순임의추출, 층화추출, 계통추출, 집락추출 중 어느 표본추출방법에 해당하는지 구분하는 문제이다. 모집단을 어떤 기준으로 나누고, 나눈 각 부분에서 표본을 어떻게 뽑았는지를 확인하면 방법이 갈린다.

핵심 개념

층화추출법은 모집단을 서로 성질이 다른 하위 집단(층)으로 나눈 뒤, 각 층에서 임의로 표본을 뽑는 방법이다. 층 내부는 동질적이고 층 사이에는 이질적이도록 나누는 것이 특징이며, 각 층에서 반드시 표본이 추출되므로 특정 층이 표본에서 누락되는 일이 없다.
집락(군집)추출법은 모집단을 여러 집락으로 나눈 뒤, 집락 중 일부를 통째로 무작위로 선택하고 선택된 집락 안의 모든 개체(또는 그 안에서 다시 추출)를 조사하는 방법이다. 층화추출과 달리 선택되지 않은 집락은 아예 조사 대상에서 빠질 수 있다.
단순임의추출법은 모집단 전체에서 별도의 구분 없이 동일한 확률로 표본을 뽑는 방법이고, 계통추출법은 모집단에 일련번호를 매긴 뒤 일정한 간격으로 표본을 뽑는 방법이다.

NoteBase

선지별 해설

① 단순임의추출법(simple random sampling) — X. 서울특별시 전체 가구를 하나의 집단으로 놓고 무작위로 뽑은 것이 아니라, 25개 구로 미리 나눈 뒤 각 구에서 따로 표본을 뽑았으므로 단순임의추출이 아니다.
② 층화추출법(stratified sampling) — O. 서울특별시를 25개 구라는 하위 집단(층)으로 나누고, 25개 층 모두에서 빠짐없이 각각 20개씩 가구를 임의로 뽑았다. 모든 층에서 표본이 추출되고 각 층의 표본 크기가 동일하게 배정된 전형적인 층화추출법(그중에서도 동일 배분) 사례이다.
③ 계통추출법(systematic sampling) — X. 가구에 일련번호를 매겨 일정한 간격으로 뽑았다는 내용이 없으므로 계통추출이 아니다.
④ 집락(군집)추출법(cluster sampling) — X. 25개 구 전부가 조사 대상에 포함되어 있고 각 구에서 다시 20개씩 임의로 뽑았다. 집락추출이라면 25개 구 중 일부 구만 무작위로 선택하고 선택된 구는 통째로(또는 그 안에서만) 조사해야 하는데, 이 사례는 구를 선택하는 단계 없이 25개 구 전부에서 표본을 뽑았으므로 집락추출이 아니다.

오답 포인트

집락추출법과 층화추출법을 혼동하기 쉬운데, 핵심은 '하위 집단을 전부 조사에 포함시키는가, 아니면 일부만 무작위로 골라 그 안을 조사하는가'이다. 이 사례는 25개 구 전부에서 표본을 뽑았으므로 층화추출이며, 만약 25개 구 중 예를 들어 5개 구만 무작위로 골라 그 구의 가구를 조사했다면 집락추출이 된다. 또한 '임의로 뽑아'라는 표현만 보고 단순임의추출로 오인하지 않도록, 그 앞에 있는 '25개 각 구로부터'라는 층화 구조를 놓치지 않아야 한다.

다시 나오면 이렇게 판단

먼저 모집단이 몇 개의 하위 집단으로 나뉘어 있는지 확인하고, 그 하위 집단 전부에서 표본이 나오는지 일부에서만 나오는지를 본다. 전부에서 나오면 층화추출, 일부 하위 집단만 골라 그 안을 통째로(또는 표본을) 조사하면 집락추출, 하위 집단 구분 없이 전체에서 무작위로 뽑으면 단순임의추출, 번호를 매겨 일정 간격으로 뽑으면 계통추출로 판단한다.

암기 포인트

층화추출=모든 층에서 뽑는다(층 내 동질, 층 간 이질), 집락추출=일부 집락만 뽑아 그 안을 조사한다(집락 내 이질, 집락 간 동질). '25개 구 전부에서 20개씩'처럼 하위 집단 전체가 등장하면 층화추출을 먼저 떠올린다.

NoteBase

문 2. 다음은 기초통계학을 수강한 학생 20명의 성적을 줄기-잎 그림으로 나타낸 것이다. 이에 대한 설명으로 옳지 않은 것은?

5	2
6	34778
7	23466668
8	23446
9	0

① 범위는 38이다.

② 중앙값은 76이다.

③ 최빈값은 76이다.

④ 제1 사분위수는 82.5이다.

정답 4번

출제 포인트

줄기-잎 그림에서 원자료를 복원한 뒤 범위·중앙값·최빈값·사분위수 같은 기술통계량을 정확히 계산할 수 있는지를 묻는다. 특히 사분위수는 중앙값을 기준으로 자료를 상위·하위 절반으로 나눈 뒤 각 절반의 중앙값으로 구한다는 점을 이해해야 하며, 이 문제는 제3사분위수 값을 제1사분위수라고 잘못 제시한 오류를 찾는 것이 핵심이다.

조건 정리

줄기-잎 그림에서 줄기는 십의 자리, 잎은 일의 자리를 나타낸다. 이를 풀어 정리하면 학생 20명의 성적은 다음과 같다.
52
63, 64, 67, 67, 68
72, 73, 74, 76, 76, 76, 76, 78
82, 83, 84, 84, 86
90
이를 작은 값부터 순서대로 나열하면 52, 63, 64, 67, 67, 68, 72, 73, 74, 76, 76, 76, 76, 78, 82, 83, 84, 84, 86, 90이며 총 20개(짝수 개)이다.

NoteBase

사용할 공식

범위 = 최댓값 − 최솟값.
중앙값(짝수 개, n=20) = (10번째 값 + 11번째 값) / 2.
최빈값 = 가장 많이 나타나는 값.
사분위수는 중앙값으로 자료를 하위 10개와 상위 10개로 나눈 뒤, 하위 10개의 중앙값을 제1사분위수(Q_1), 상위 10개의 중앙값을 제3사분위수(Q_3)로 계산한다.

계산 과정

범위: 최댓값 90, 최솟값 52이므로 범위 = 90 − 52 = 38이다. ①은 옳은 설명이다.
중앙값: 정렬된 20개 값 중 10번째 값은 76, 11번째 값도 76이므로 중앙값 = (76+76)/2 = 76이다. ②는 옳은 설명이다.
최빈값: 76이 네 번(10~13번째) 나타나 가장 많이 등장하므로 최빈값은 76이다. ③은 옳은 설명이다.
사분위수: 하위 10개는 52, 63, 64, 67, 67, 68, 72, 73, 74, 76이고 이 중 5번째(67)와 6번째(68)의 평균이 제1사분위수이므로 $Q_1 = (67 + 68)/2 = 67.5$이다. 반면 상위 10개는 76, 76, 76, 78, 82, 83, 84, 84, 86, 90이고 이 중 5번째(82)와 6번째(83)의 평균이 제3사분위수이므로 $Q_3 = (82 + 83)/2 = 82.5$이다. 즉 82.5는 제1사분위수가 아니라 제3사분위수 값이다.

선지 판단

① 범위는 38이다 → O. 90−52=38로 계산과 일치한다.
② 중앙값은 76이다 → O. 10번째·11번째 값이 모두 76이므로 중앙값은 76이다.
③ 최빈값은 76이다 → O. 76이 네 번으로 가장 많이 나타난다.
④ 제1 사분위수는 82.5이다 → X. 82.5는 실제로는 제3사분위수(Q_3) 값이고, 제1사분위수(Q_1)는 67.5이다. 따라서 이 설명은 옳지 않다.

오답 포인트

①②③을 모두 정확한 값으로 확인하고도 ④를 무심코 넘기면 실수하기 쉽다. 사분위수를 구할 때 하위 절반과 상위 절반을 뒤바꿔 계산하거나, 제1사분위수와 제3사분위수의 값을 헷갈려 서로 바꿔 쓰는 것이 이 문제가 노리는 함정이다. 또한 중앙값 계산 시 20개(짝수 개)이므로 가운데 두 값의 평균을 내야 하는데, 한 값만 취하는 실수도 주의해야 한다.

다시 나오면 이렇게 판단

줄기-잎 그림 문제는 먼저 원자료를 오름차순으로 완전히 복원한 다음 개수를 세어 홀수/짝수 개를 확인한다. 사분위수를 구할 때는 전체 중앙값으로 자료를 정확히 반으로 나눈 뒤, 하위 절반의 중앙값은 제1사분위수, 상위 절반의 중앙값은 제3사분위수라는 것을 순서대로 적용한다. 선지에 나온 숫자가 어느 사분위수에 해당하는지 반드시 직접 계산해 대조한다.

NoteBase

암기 포인트

제1사분위수(Q_1)=하위 절반의 중앙값, 제3사분위수(Q_3)=상위 절반의 중앙값이라는 점을 짝지어 기억해 둔다. 범위=최댓값−최솟값, 짝수 개 자료의 중앙값=가운데 두 값의 평균, 최빈값=가장 많이 나온 값이라는 기본 정의도 함께 정리해 둔다.

NoteBase

문 3. 관측값 $x_1, x_2, \cdots, x_n$의 표본평균을 $\bar{x}$, 표본분산을 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$으로 정의할 때, $n = 10$인 경우 $\frac{10\sum_{i=1}^{10} x_i^2 - \left(\sum_{i=1}^{10} x_i\right)^2}{s^2}$의 값은? (단, $x_1, x_2, \cdots, x_{10}$의 값이 모두 같지는 않다)

① 9

② 10

③ 90

④ 100

정답 3번

출제 포인트

표본분산 s^2의 정의식과 편차제곱합의 계산 공식 $\sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$ 사이의 관계를 이용해 주어진 식을 n과 s^2만의 관계식으로 정리할 수 있는지를 묻는다.

조건 정리

$n = 10$이고 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$이다. 구해야 할 값은 $\frac{10\sum_{i=1}^{10} x_i^2 - \left(\sum_{i=1}^{10} x_i\right)^2}{s^2}$이다. $x_1, \cdots, x_{10}$이 모두 같지 않다는 조건은 $s^2 \neq 0$임을 보장해 나눗셈이 가능하도록 하는 조건이다.

사용할 공식

편차제곱합의 전개 공식 $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$을 이용한다. 이 식의 양변에 n을 곱하면 $n\sum x_i^2 - (\sum x_i)^2 = n\sum(x_i - \overline{x})^2$이 되고, 표본분산의 정의 $s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$로부터 $\sum(x_i - \overline{x})^2 = (n-1)s^2$이다.

계산 과정

위 두 관계를 결합하면 $n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = n(n-1)s^2$이 성립한다. 따라서 $\frac{n\sum x_i^2 - (\sum x_i)^2}{s^2} = n(n-1)$이다. $n = 10$을 대입하면 $10 \times (10-1) = 10 \times 9 = 90$이다.

선지 판단

① 9 — X. $n - 1$만 계산한 값으로, n을 곱하지 않아 틀린 값이다.

② 10 — X. n 값 자체이며, 식의 값과는 무관하다.

③ 90 — O. $n(n-1) = 10 \times 9 = 90$으로 계산 결과와 일치한다.

④ 100 — X. $n^2 = 10 \times 10$을 계산한 값으로, $n - 1$ 대신 n을 두 번 곱한 오류에 해당한다.

NoteBase

오답 포인트

이 문제의 함정은 $n\sum x_i^2 - (\sum x_i)^2$을 $\sum(x_i - \overline{x})^2$과 바로 같다고 착각하는 데 있다. 실제로는 n배 관계, 즉 $n\sum x_i^2 - (\sum x_i)^2 = n\sum(x_i - \overline{x})^2$이므로 최종적으로 s^2으로 나누면 $n-1$이 아니라 $n(n-1)$이 남는다는 점을 놓치면 9번이나 100번을 고르게 된다.

다시 나오면 이렇게 판단

표본분산이나 편차제곱합이 포함된 식이 나오면 우선 $\sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$ 공식으로 치환한 뒤, 남은 식을 n과 $(n-1)$로 정리하는 순서로 접근한다. 계수에 곱해진 n을 빠뜨리지 않도록 양변에 곱해지는 상수를 끝까지 따라간다.

암기 포인트

편차제곱합 전개 공식 $\sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$과, 이를 변형한 $n\sum x_i^2 - (\sum x_i)^2 = n(n-1)s^2$의 관계를 함께 기억해 둔다.

NoteBase

문 4. 자유도가 5인 카이제곱분포에 대한 설명으로 옳은 것은?

① 기댓값은 중앙값보다 크다.

② 왜도(skewness)는 음수이다.

③ 기댓값은 4이다.

④ 분산은 5이다.

정답 1번

출제 포인트

자유도 k인 카이제곱분포 $\chi^2(k)$의 기댓값·분산·왜도·중앙값을 정확히 계산해 각 선지의 수치·부등호 진술을 검증할 수 있는지를 묻는다. 카이제곱분포는 오른쪽으로 꼬리가 긴(양의 왜도) 비대칭 분포라는 점이 선지 판단의 핵심이다.

조건 정리

자유도 $k = 5$인 카이제곱분포 $X \sim \chi^2(5)$를 가정한다. 카이제곱분포는 감마분포의 특수한 경우로, k가 커질수록 정규분포에 가까워지지만 $k = 5$처럼 자유도가 작을 때는 오른쪽으로 치우친 비대칭 형태를 뚜렷이 보인다.

사용할 공식

카이제곱분포 $\chi^2(k)$의 성질은 다음과 같다. 기댓값 $E(X) = k$, 분산 $Var(X) = 2k$, 왜도 $\gamma_1 = \sqrt{8/k}$(항상 양수), 중앙값은 근사적으로 $k\left(1 - \frac{2}{9k}\right)^3$으로 기댓값보다 작다.

계산 과정

$k = 5$를 대입하면 기댓값은 $E(X) = 5$이고 분산은 $Var(X) = 2 \times 5 = 10$이다. 왜도는 $\gamma_1 = \sqrt{8/5} = \sqrt{1.6} \approx 1.26$으로 0보다 큰 양수이며, 이는 분포가 오른쪽 꼬리가 긴 형태임을 뜻한다. 중앙값을 근사식으로 구하면 $5\left(1 - \frac{2}{45}\right)^3 \approx 5 \times 0.878 \approx 4.39$로, 기댓값 5보다 작다. 따라서 기댓값(5)이 중앙값(약 4.39)보다 크다는 것이 성립한다.

선지 판단

① 기댓값(5)은 중앙값(약 4.39)보다 크다 → 계산 결과와 일치하므로 옳다.
② 왜도는 $\sqrt{8/5} \approx 1.26 > 0$으로 양수이다 → 틀린 표현(음수이다) → 올바른 표현(왜도는 양수이며 분포는 오른쪽으로 치우쳐 있다).
③ 기댓값은 $E(X) = k = 5$이다 → 틀린 표현(기댓값은 4이다) → 올바른 표현(기댓값은 5이다).
④ 분산은 $Var(X) = 2k = 10$이다 → 틀린 표현(분산은 5이다) → 올바른 표현(분산은 10이다).

NoteBase

오답 포인트

카이제곱분포의 기댓값(k)과 분산($2k$) 공식을 헷갈려 분산 자리에 기댓값 값을, 기댓값 자리에 다른 수치를 넣는 실수가 흔하다. 또한 카이제곱분포는 항상 오른쪽으로 긴 꼬리를 가지는 양의 왜도 분포라는 점을 몰라 왜도의 부호를 반대로 답하기 쉽다.

다시 나오면 이렇게 판단

자유도 k가 주어지면 가장 먼저 기댓값 k, 분산 $2k$를 계산해 선지의 숫자와 바로 대조한다. 왜도·비대칭성을 묻는 선지가 있으면 카이제곱분포는 항상 오른쪽 꼬리가 긴 양의 왜도를 가지며, 그 결과 기댓값이 중앙값보다 크다는 대소관계를 기억해 둔다.

암기 포인트

카이제곱분포 $\chi^2(k)$: 기댓값 $=k$, 분산 $=2k$, 왜도 $=\sqrt{8/k}>0$(항상 양의 왜도, 오른쪽 꼬리). 항상 기댓값 > 중앙값 > 최빈값($=k-2, k\geq 2$)의 순서로 정렬된다.

NoteBase

문 5. 수명이 평균 100일인 지수분포를 따르는 트랜지스터를 100일 동안 사용했을 때, 앞으로 100일 더 사용할 수 있을 확률은? (단, 평균이 μ인 지수분포의 확률밀도함수는 다음과 같다)

f(x)=cases(1)/(μ)e^-(1)/(μ)x, & x≥ 0 \ 0, & 그 외cases(μ>0)

① $e^{-1/2}$

② e^{-1}

③ e^{-2}

④ e^{-3}

정답 2번

출제 포인트

지수분포의 무기억성(memoryless property)을 이용해 조건부 확률 $P(X > 200 \mid X > 100)$을 구하는 문제이다. 이미 100일을 사용한 상태에서 다시 100일을 더 쓸 확률을 구하라는 조건이 핵심이다.

조건 정리

트랜지스터의 수명 X는 평균 $\mu = 100$인 지수분포를 따른다. 100일 동안 이미 사용했다는 조건 하에서 앞으로 100일을 더 사용할 확률, 즉 $P(X > 200 \mid X > 100)$을 구해야 한다.

사용할 공식

지수분포는 무기억성을 가지므로 임의의 $s, t \geq 0$에 대해 $P(X > s + t \mid X > s) = P(X > t)$가 성립한다. 또한 지수분포의 생존함수는 $P(X > x) = e^{-\frac{1}{\mu}x}$이다.

계산 과정

무기억성에 의해 $P(X > 200 \mid X > 100) = P(X > 100)$이다. $\mu = 100$이므로
$P(X > 100) = e^{-\frac{1}{100} \cdot 100} = e^{-1}$이 된다. 직접 조건부 확률의 정의로 계산해도 같은 결과를 얻는다:
$P(X > 200 \mid X > 100) = \frac{P(X > 200)}{P(X > 100)} = \frac{e^{-200/100}}{e^{-100/100}} = \frac{e^{-2}}{e^{-1}} = e^{-1}$.

선지 판단

① $e^{-1/2}$: X. $\mu = 100$에서 100일을 더 쓸 확률은 지수부가 -1이 되어야 하므로 틀린 값이다.
② e^{-1}: O. 무기억성과 조건부 확률 정의 계산 모두 e^{-1}로 일치한다.
③ e^{-2}: X. 이는 조건 없이 $P(X > 200)$만 계산한 값과 같다. 조건부 확률로 나누는 과정을 빠뜨리면 나오는 오답이다.
④ e^{-3}: X. 100일과 200일을 단순히 더한 300일에 대한 생존확률 $P(X > 300) = e^{-3}$에 해당하는 값으로, 문제에서 묻는 조건부 확률이 아니다.

NoteBase

오답 포인트

조건부 확률을 무시하고 $P(X > 200)$만 계산하면 ③이 나오고, 100일과 200일을 합쳐 300일에 대한 확률을 구하면 ④가 나온다. 지수분포는 무기억성 때문에 이미 사용한 100일이 앞으로의 확률에 아무 영향을 주지 않는다는 점을 놓치면 이런 오답에 빠진다.

다시 나오면 이렇게 판단

지수분포에서 '이미 t만큼 사용했을 때 앞으로 s만큼 더 사용할 확률'을 묻는 문제는 항상 무기억성을 먼저 떠올린다. 답은 조건과 무관하게 $P(X > s) = e^{-s/\mu}$로 바로 구해지며, 굳이 조건부 확률 공식으로 분자·분모를 나눠 계산해도 결과는 같다.

암기 포인트

지수분포의 무기억성: $P(X > s + t \mid X > s) = P(X > t)$. 지수분포의 생존함수: $P(X > x) = e^{-x/\mu}$. 이 문제처럼 '추가로 얼마나 더'를 묻는 조건부 확률은 추가 기간만 대입하면 바로 답이 나온다.

NoteBase

문 6. 어느 공장에서 생산하는 볼트의 지름(단위: mm)은 $N(14, 0.3^2)$, 너트의 지름(단위: mm)은 $N(15, 0.4^2)$을 따르고 볼트의 지름과 너트의 지름은 서로 독립이라고 하자. 이 공장에서 볼트와 너트를 임의로 하나씩 선택할 때, 볼트가 너트에 들어가지 않을 확률은? (단, $\Phi(x)$는 표준정규분포의 누적분포함수이다)

① $\Phi(-2)$

② $\Phi(2)$

③ $\Phi(-1)$

④ $\Phi(1)$

정답 1번

출제 포인트

서로 독립인 두 정규분포 확률변수의 차 $X-Y$도 정규분포를 따른다는 사실을 이용해, 두 확률변수의 대소 비교 확률을 표준정규분포 누적분포함수 Φ로 표현하는 문제이다. 볼트가 너트에 들어가려면 볼트 지름이 너트 지름보다 작아야 하므로, '들어가지 않을 확률'은 볼트 지름이 너트 지름보다 크거나 같을 확률로 해석해야 한다.

조건 정리

볼트의 지름을 X, 너트의 지름을 Y라 하면 $X \sim N(14, 0.3^2)$, $Y \sim N(15, 0.4^2)$이고 X와 Y는 서로 독립이다. 볼트가 너트에 들어가지 않는다는 것은 볼트의 지름이 너트의 지름보다 작지 않다는 뜻이므로, 구하는 확률은 $P(X \geq Y)$, 즉 $P(X - Y \geq 0)$이다.

사용할 공식

독립인 두 정규분포 확률변수 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$에 대해 $D = X - Y$는 $D \sim N(\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2)$를 따른다. 이후 표준화 $Z = \frac{D - \mu_D}{\sigma_D}$를 이용해 $P(D \geq 0)$을 Φ로 나타낸다.

계산 과정

$D = X - Y$의 평균은 $\mu_D = 14 - 15 = -1$이고, 분산은 $\sigma_D^2 = 0.3^2 + 0.4^2 = 0.09 + 0.16 = 0.25$이므로 $\sigma_D = 0.5$이다. 따라서 $D \sim N(-1, 0.5^2)$이다.
구하는 확률은 $P(D \geq 0) = P\left(Z \geq \frac{0 - (-1)}{0.5}\right) = P(Z \geq 2) = 1 - \Phi(2) = \Phi(-2)$이다(표준정규분포의 대칭성 $1 - \Phi(z) = \Phi(-z)$ 이용).

NoteBase

선지 판단

① Φ(-2): 계산 결과와 일치한다. $P(X \geq Y) = P(Z \geq 2) = \Phi(-2)$이므로 옳다.
② Φ(2): $\Phi(2)$는 $P(Z \leq 2)$에 해당하는 값으로, 구하는 확률인 $P(Z \geq 2)$와는 여사건 관계(합이 1이 아니라 $\Phi(2) + \Phi(-2) = 1$)이므로 틀린 표현이다 → 올바른 표현은 Φ(-2).
③ Φ(-1): 표준화 값을 $z = 1$이 아닌 $z = 1$로 잘못 계산하거나 평균 또는 분산을 잘못 대입했을 때 나오는 값으로, 실제 표준화 값은 $z = 2$이므로 틀린 표현이다 → 올바른 표현은 Φ(-2).
④ Φ(1): 분산을 더하지 않고 표준편차를 단순히 빼는 등 계산 과정에서 오차가 생겼을 때 나올 수 있는 값으로, 옳지 않다 → 올바른 표현은 Φ(-2).

오답 포인트

이 문제에서 가장 흔한 실수는 두 가지이다. 첫째, '볼트가 너트에 들어가지 않는다'는 조건을 $P(X \leq Y)$로 잘못 세우는 경우이다(부등호 방향 반대). 둘째, 분산을 더할 때 표준편차를 그대로 더하거나 빼는 경우이다(분산은 더하되 표준편차는 제곱합의 제곱근으로 구해야 한다). 또한 $\Phi(2)$와 $\Phi(-2)$를 혼동하지 않도록 표준정규분포의 대칭성 $1 - \Phi(z) = \Phi(-z)$를 정확히 적용해야 한다.

다시 나오면 이렇게 판단

독립인 두 정규분포 확률변수의 대소 비교 확률을 묻는 문제는, 먼저 두 변수의 차를 새로운 확률변수로 정의하고(예: $D = X - Y$) 그 평균과 분산을 각각 구한 뒤 표준화하여 Φ로 표현하는 순서로 접근한다. 이때 '~하지 않을 확률', '~보다 큰 경우' 등 문장에서 부등호 방향을 정확히 확정하는 것이 핵심이다.

암기 포인트

독립인 정규분포 확률변수의 차 $D = X - Y \sim N(\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2)$: 평균은 빼고 분산은 더한다.
표준정규분포의 대칭성 $1 - \Phi(z) = \Phi(-z)$도 함께 기억해 둔다.

NoteBase

문 7. 평균과 분산이 알려지지 않은 정규모집단으로부터 16개의 표본을 임의추출한 결과 표본평균 $\overline{x}=13$, 표본분산 $s^2=\frac{1}{15}\sum_{i=1}^{16}(x_i-\overline{x})^2=9$를 얻었다. 모평균 μ에 대한 가설 $H_0:\mu=15$ 대 $H_1:\mu\neq 15$를 검정할 때, p-값은? (단, Z와 T는 각각 표준정규분포와 자유도가 15인 t-분포를 따르는 확률변수이다)

① $2P\left(Z\geq\frac{2\sqrt{15}}{3}\right)$
② $2P\left(Z\geq\frac{8}{3}\right)$
③ $2P\left(T\geq\frac{2\sqrt{15}}{3}\right)$
④ $2P\left(T\geq\frac{8}{3}\right)$

정답 4번

출제 포인트

모분산을 모르는 정규모집단에서 평균에 대한 양측검정을 수행할 때, 검정통계량을 t-분포를 따르는 형태로 세우고 그로부터 p-값 표현식을 만들 수 있는지를 묻는다. 모분산이 미지이고 표본분산으로 대체해야 하므로 Z가 아니라 T를 사용해야 한다는 점이 핵심이다.

조건 정리

표본크기는 $n=16$, 표본평균은 $\overline{x}=13$, 표본분산은 $s^2=\frac{1}{15}\sum_{i=1}^{16}(x_i-\overline{x})^2=9$이다(자유도 $n-1=15$로 나눈 불편분산이므로 $s=3$). 귀무가설은 $H_0:\mu=15$, 대립가설은 $H_1:\mu\neq 15$로 양측검정이다. 모분산이 알려져 있지 않으므로 모표준편차 대신 표본표준편차 s를 사용한 t-검정통계량을 쓴다.

사용할 공식

모분산 미지일 때 모평균에 대한 검정통계량은 $T=\frac{\overline{x}-\mu_0}{s/\sqrt{n}}$이며, 이는 자유도 $n-1$인 t-분포를 따른다. 양측검정의 p-값은 $p=2P(T\geq|T_{obs}|)$로 구한다.

계산 과정

표본표준편차는 $s=\sqrt{9}=3$이다. 검정통계량은 $T_{obs}=\frac{\overline{x}-\mu_0}{s/\sqrt{n}}=\frac{13-15}{3/\sqrt{16}}=\frac{-2}{3/4}=-\frac{8}{3}$이다. 자유도는 $n-1=15$이다. 양측검정이므로 p-값은 관측된 검정통계량의 절댓값을 이용해 $p=2P\left(T\geq\frac{8}{3}\right)$로 표현된다.

선지 판단

① 2P(Z≥2√15/3) → X. 모분산을 몰라 Z가 아닌 T를 써야 하고, 수치도 2√15/3이 아니라 8/3이 되어야 한다.
② 2P(Z≥8/3) → X. 수치 8/3은 맞지만 모분산 미지 상황에서는 Z가 아니라 T를 써야 한다.
③ 2P(T≥2√15/3) → X. T를 쓴 것은 맞지만 검정통계량 값이 2√15/3이 아니라 8/3이다(√n 처리 오류에 해당하는 값).
④ 2P(T≥8/3) → O. 모분산 미지이므로 T를 사용하고, 계산된 검정통계량의 절댓값이 8/3이므로 p-값은 2P(T≥8/3)이다.

NoteBase

오답 포인트

가장 흔한 실수는 모분산을 모른다는 조건을 놓치고 Z-검정 공식을 그대로 쓰는 것이다(①, ②). 또한 검정통계량을 계산할 때 $s/\sqrt{n}$을 $\sqrt{s^2/n}$ 형태로 잘못 다뤄 $2\sqrt{15}/3$ 같은 엉뚱한 값을 만드는 실수도 나온다(①, ③). $s^2 = 9$이므로 $s = 3$이고 $\sqrt{n} = 4$임을 정확히 계산해야 한다.

다시 나오면 이렇게 판단

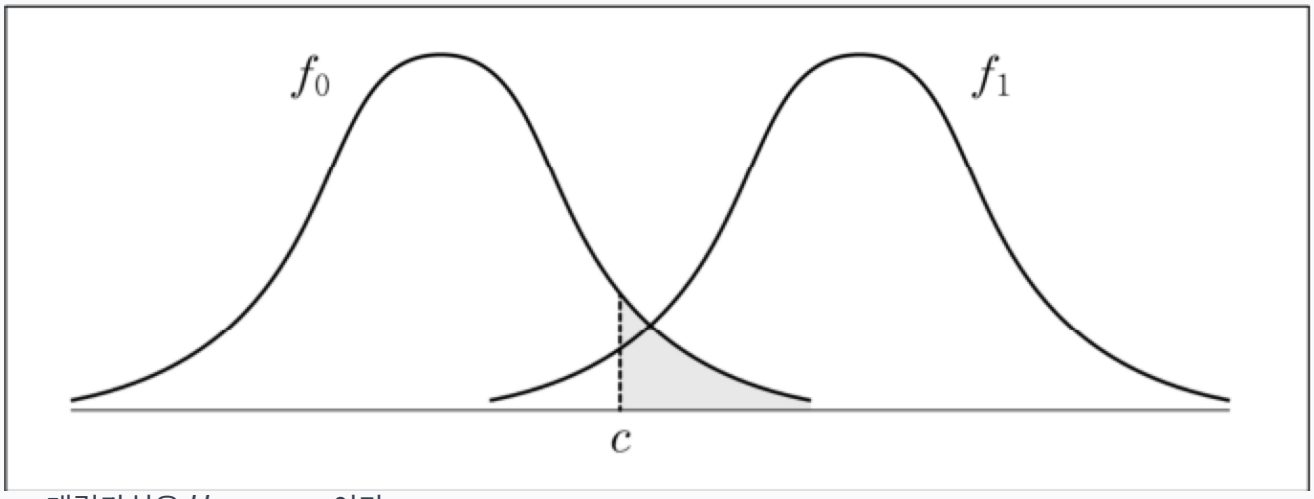
문제에서 모분산이 알려지지 않았다는 표현이 나오면 곧바로 T-검정통계량을 떠올린다. 이후 s, n, $\overline{x} - \mu_0$을 각각 계산해 $T = (\overline{x} - \mu_0)/(s/\sqrt{n})$에 대입하고, 양측검정이므로 절댓값을 취해 $2P(T \geq |T_{obs}|)$ 형태로 정리하면 된다.

암기 포인트

모분산 기지 → Z-검정, 모분산 미지 → T-검정(자유도 n-1). 양측검정 p-값 = 2P(· ≥ |검정통계량|).
표본표준편차는 표본분산의 제곱근이다.

NoteBase

문 8. 다음은 가설 $H_0 : \mu = \mu_0$에 대한 유의수준 α인 단측검정에 관한 그림이다. 그림에서 f_0와 f_1으로 표시된 분포는 각각 귀무가설과 대립가설의 특정 모숫값하에서의 검정통계량의 분포를 나타내며, c는 검정의 임계값(critical value)을 나타낸다. 이에 대한 설명으로 옳은 것만을 모두 고른 것은?



ㄱ. 대립가설은 $H_1 : \mu < \mu_0$이다.

ㄴ. 음영 부분의 넓이는 α이다.

ㄷ. f_1이 우측으로 이동할수록 검정력은 커진다.

ㄹ. c를 검정통계량의 값으로 정할 때 음영부분의 넓이가 p-값이다.

① ㄱ, ㄹ

② ㄴ, ㄷ

③ ㄴ, ㄹ

④ ㄴ, ㄷ, ㄹ

정답 4번

출제 포인트

귀무가설과 대립가설 하의 검정통계량 분포(f0, f1)와 임곗값 c의 위치 관계를 그림으로 제시하고, 이로부터 대립가설의 방향, 유의수준 α, 검정력, p-값의 개념을 정확히 구분할 수 있는지를 묻는 문제이다.

NoteBase

핵심 개념

f0는 귀무가설 $H_0 : \mu = \mu_0$ 하에서의 검정통계량 분포이고, f1은 대립가설의 한 특정 모숫값 하에서의 검정통계량 분포이다. 그림에서 f1이 f0보다 오른쪽에 위치하므로, 검정통계량 값이 클수록 대립가설 쪽 분포에 가깝다는 뜻이 되어 대립가설은 모평균이 μ0보다 크다는 방향이 된다. 기각역은 c보다 큰 쪽(오른쪽 꼬리)이다. 유의수준 α는 귀무가설이 참인데도 기각하게 될 확률이므로 f0 곡선에서 c 오른쪽 부분의 넓이이고, 검정력은 대립가설이 참일 때 올바르게 기각할 확률이므로 f1 곡선에서 c 오른쪽 부분의 넓이이다. p-값은 실제로 관측된 검정통계량 값을 기준으로, 귀무가설 하에서 그보다 더 극단적인(이 경우 더 큰) 값이 나올 확률이다.

선지별 해설

ㄱ. 그림에서 f1은 f0보다 오른쪽에 위치한다. 이는 검정통계량이 클수록 대립가설 쪽 분포에 가까운 값이라는 의미이므로, 대립가설은 모평균이 μ0보다 크다는 방향, 즉 $H_1 : \mu > \mu_0$이 되어야 한다. "$H_1 : \mu < \mu_0$"이라는 설명은 틀렸다.
ㄴ. 음영 부분은 c보다 오른쪽에서 f0(귀무가설 하의 분포) 곡선 아래에 해당하는 넓이이다. 이는 귀무가설이 참인데도 검정통계량이 c보다 커서 귀무가설을 기각하게 될 확률과 같으므로 유의수준 α이다. 옳은 설명이다.
ㄷ. 검정력은 대립가설이 참일 때 귀무가설을 옳게 기각할 확률, 즉 f1 곡선에서 c보다 오른쪽 부분의 넓이이다. f1이 우측으로 이동하면 c보다 오른쪽에 놓이는 f1의 넓이가 늘어나므로 검정력은 커진다. 옳은 설명이다.
ㄹ. p-값은 실제로 관측된 검정통계량 값을 기준으로, 귀무가설 하에서 그보다 더 큰 값이 나올 확률이다. 그림에서 c를 미리 정해둔 고정 임곗값이 아니라 실제 관측된 검정통계량의 값으로 놓으면, 그 오른쪽에서 f0 곡선 아래의 넓이가 바로 p-값이 된다. 옳은 설명이다.
따라서 옳은 것은 ㄴ, ㄷ, ㄹ이다.
① ㄱ, ㄹ: ㄱ이 틀린 설명이므로 옳지 않다.
② ㄴ, ㄷ: 옳은 설명인 ㄹ이 빠져 있으므로 옳지 않다.
③ ㄴ, ㄹ: 옳은 설명인 ㄷ이 빠져 있으므로 옳지 않다.
④ ㄴ, ㄷ, ㄹ: 옳은 설명을 모두 포함하므로 정답이다.

오답 포인트

f1이 f0보다 오른쪽에 그려져 있다는 것만 보고 대립가설의 부등호 방향을 습관적으로 반대로 판단하는 실수를 하기 쉽다. 분포가 오른쪽에 있다는 것은 검정통계량이 클수록 그 분포에 가깝다는 뜻이므로 대립가설도 '크다' 방향이 되어야 함을 놓치지 않아야 한다. 또한 α와 p-값을 혼동하기 쉬운데, 두 값은 계산 방식이 같고(둘 다 f0의 오른쪽 꼬리 넓이) 차이는 경계값이 사전에 정한 고정 임곗값 c인지, 실제 관측된 검정통계량 값인지에만 있다.

다시 나오면 이렇게 판단

이런 유형은 먼저 f0(귀무)와 f1(대립) 중 어느 쪽이 더 오른쪽에 있는지를 확인해 대립가설의 부등호 방향을 정한다. 다음으로 음영 부분이 어느 분포의 어느 꼬리 넓이인지를 확인해 α(f0의 기각역 넓이)인지 검정력(f1의 기각역 넓이)인지를 구분한다. 마지막으로 경계값이 사전에 정한 고정 임곗값 c인지, 실제 관측된 통계량 값인지에 따라 그 넓이가 α인지 p-값인지를 판단한다.

NoteBase

암기 포인트

α = 귀무가설 분포(f0)에서 고정 임곗값 기준 기각역 넓이. 검정력 = 대립가설 분포(f1)에서 고정 임곗값 기준 기각역 넓이. p-값 = 귀무가설 분포(f0)에서 관측된 통계량 값 기준 기각역 넓이.

NoteBase

문 9. 표준정규분포와 자유도가 r인 t-분포에 대한 설명으로 옳지 않은 것은? (단, $r > 2$)

① 두 분포 모두 평균이 0이고, 평균에 대하여 대칭이다.

② t-분포의 분산은 1보다 크다.

③ t-분포의 제95 백분위수는 표준정규분포의 제95 백분위수보다 작다.

④ 자유도 r이 무한대로 커지면 t-분포는 표준정규분포로 수렴한다.

정답 3번

출제 포인트

표준정규분포와 자유도 r인 t-분포의 공통점과 차이점을 정확히 알고 있는지 확인하는 문제다. 두 분포 모두 0을 중심으로 좌우 대칭인 종 모양 분포이지만, t-분포는 자유도가 유한할 때 표준정규분포보다 꼬리가 두꺼운(퍼짐이 큰) 분포라는 점이 핵심이다.

핵심 개념

표준정규분포 Z는 평균 0, 분산 1인 분포다. 자유도 r인 t-분포도 평균은 0이지만 분산은 $\frac{r}{r-2}$ (단, $r > 2$)로 항상 1보다 크다. 분산이 더 크다는 것은 분포가 표준정규분포보다 옆으로 더 퍼져 있고 꼬리가 두껍다는 뜻이다. 꼬리가 두꺼우면 같은 누적확률(백분위수)에 도달하기 위해 표준정규분포보다 더 먼 값이 필요하므로, t-분포의 상위 백분위수(예: 제95 백분위수)는 표준정규분포의 대응 백분위수보다 항상 크다. 또한 자유도 r이 커질수록 t-분포의 분산 $\frac{r}{r-2}$은 1에 가까워지고, 분포 형태 자체가 표준정규분포로 수렴한다.

선지별 해설

① 두 분포 모두 평균이 0이고, 평균에 대하여 대칭이다. → 옳다(O). 표준정규분포와 t-분포 모두 0을 중심으로 좌우 대칭인 종 모양 분포다.

② t-분포의 분산은 1보다 크다. → 옳다(O). $r > 2$일 때 t-분포의 분산은 $\frac{r}{r-2}$이며, 이 값은 항상 1보다 크다(예: $r = 10$이면 $\frac{10}{8} = 1.25$).

③ t-분포의 제95 백분위수는 표준정규분포의 제95 백분위수보다 작다. → 옳지 않다(X). t-분포는 표준정규분포보다 꼬리가 두꺼워 퍼짐이 크므로, 같은 누적확률 95%에 도달하는 값은 오히려 표준정규분포보다 크다. 즉 t-분포의 제95 백분위수는 표준정규분포의 제95 백분위수보다 작지 않고 더 크다. 이 선지가 틀린 설명이므로 정답이다.

④ 자유도 r이 무한대로 커지면 t-분포는 표준정규분포로 수렴한다. → 옳다(O). $r \to \infty$이면 t-분포의 분산 $\frac{r}{r-2} \to 1$이 되고, 분포 형태도 표준정규분포와 일치하게 된다.

NoteBase

오답 포인트

①②④는 모두 옳은 설명이라 소거되고, ③만 틀린 설명이라 정답이 된다. 흔한 실수는 '분산이 더 크다'와 '백분위수가 더 작다'를 반대로 연결하는 것이다. 분산(퍼짐)이 크면 극단값 쪽으로 더 멀리 퍼지므로, 상위 백분위수(예: 95%, 99%)는 오히려 더 커진다는 방향을 정확히 기억해야 한다.

다시 나오면 이렇게 판단

표준정규분포와 t-분포를 비교하는 문제는 '분산 크기'와 '꼬리 두께 및 상위 백분위수 크기'가 같은 방향으로 움직인다는 점만 기억하면 빠르게 판단할 수 있다. t-분포의 분산 공식 $\frac{r}{r-2}$을 떠올리고, 이 값이 1보다 큰지 자유도에 따라 어떻게 변하는지를 먼저 확인한 뒤 각 선지를 대조한다.

암기 포인트

t-분포 분산 $= \frac{r}{r-2} > 1 (r > 2)$; 꼬리 두꺼움 → 상위 백분위수는 표준정규분포보다 큼; $r \to \infty$이면 t-분포 → 표준정규분포로 수렴.

NoteBase

문 10. $X_1, X_2, \cdots, X_{10}$**을** $N(\mu, 1)$**으로부터의 확률표본이라고 하자. 가설** $H_0 : \mu = 5$ **대** $H_1 : \mu \neq 5$**를 유의수준** α**에서 검정할 때,** $\mu = 4$**에서의 검정력은? (단,** z_α**와** $\Phi(x)$**는 각각 표준정규분포의 제**$100 \times (1-\alpha)$**백분위수와 누적분포함수를 나타낸다)**

① $\Phi\left(\sqrt{10} - z_{\alpha/2}\right)$

② $1 - \Phi\left(\sqrt{10} - z_{\alpha/2}\right)$

③ $\Phi\left(\sqrt{10} + z_{\alpha/2}\right) + \Phi\left(\sqrt{10} - z_{\alpha/2}\right)$

④ $1 - \Phi\left(\sqrt{10} + z_{\alpha/2}\right) + \Phi\left(\sqrt{10} - z_{\alpha/2}\right)$

정답 4번

출제 포인트

정규모집단의 분산이 알려진 경우 모평균에 대한 양측검정의 기각역을 세우고, 참값이 귀무가설과 다를 때(μ=4) 그 기각역에서 검정력을 구하는 문제이다. 검정력은 대립가설이 참일 때 귀무가설을 기각할 확률이므로, 표본평균의 실제 분포(μ=4 기준)를 기준으로 기각역에 해당하는 확률을 계산해야 한다.

조건 정리

$X_1, \ \dots, X_{10}$은 $N(\mu, 1)$에서의 확률표본이므로 $\bar{X} \sim N(\mu, 1/10)$이다. 가설은 $H_0 : \mu = 5$ 대 $H_1 : \mu \neq 5$인 양측검정이고 유의수준은 α이다. z_α는 표준정규분포의 제 $100 \times (1-\alpha)$ 백분위수, 즉 $P(Z > z_\alpha) = \alpha$를 만족하는 값이다. 구하는 것은 참값이 $\mu = 4$일 때의 검정력이다.

사용할 공식

분산이 1로 알려진 정규모집단의 평균에 대한 양측검정에서 기각역은 $\left|\frac{\bar{X} - 5}{1/\sqrt{10}}\right| > z_{\alpha/2}$이다. 즉 $\bar{X} > 5 + \frac{z_{\alpha/2}}{\sqrt{10}}$ 또는 $\bar{X} < 5 - \frac{z_{\alpha/2}}{\sqrt{10}}$일 때 H_0를 기각한다. 검정력은 이 두 사건이 $\mu = 4$일 때 일어날 확률의 합이며, $\mu = 4$에서는 $\sqrt{10}(\bar{X} - 4) \sim N(0, 1)$임을 이용해 표준화한다.

계산 과정

먼저 $\bar{X} > 5 + \frac{z_{\alpha/2}}{\sqrt{10}}$인 확률을 $\mu = 4$ 기준으로 표준화하면
$P\left(\sqrt{10}(\bar{X} - 4) > \sqrt{10}\left(1 + \frac{z_{\alpha/2}}{\sqrt{10}}\right)\right) = P(Z > \sqrt{10} + z_{\alpha/2}) = 1 - \Phi(\sqrt{10} + z_{\alpha/2})$이다.
다음으로 $\bar{X} < 5 - \frac{z_{\alpha/2}}{\sqrt{10}}$인 확률을 같은 방식으로 표준화하면 $P\left(\sqrt{10}(\bar{X} - 4) < \sqrt{10} - z_{\alpha/2}\right) = \Phi(\sqrt{10} - z_{\alpha/2})$이다.
두 사건은 서로 배반이므로 검정력은 두 확률의 합, 즉 $1 - \Phi(\sqrt{10} + z_{\alpha/2}) + \Phi(\sqrt{10} - z_{\alpha/2})$이다.

NoteBase

선지 판단

① Φ(√10−z_{α/2}) — X . 하한 쪽(기각역 왼쪽) 확률만 남긴 식으로, 상한 쪽 기각역에서 오는 확률을 빠뜨렸다.
② 1−Φ(√10−z_{α/2}) — X . 하한 쪽 확률 Φ(√10−z_{α/2})을 오히려 여사건 형태로 잘못 적은 식이며, 상한 쪽 항도 빠져 있다.
③ Φ(√10+z_{α/2})+Φ(√10−z_{α/2}) — X . 상한 쪽 확률을 1−Φ(√10+z_{α/2}) 대신 Φ(√10+z_{α/2})로 잘못 적었다(여사건 처리 누락).
④ 1−Φ(√10+z_{α/2})+Φ(√10−z_{α/2}) — O . 상한 쪽 기각역 확률 1−Φ(√10+z_{α/2})과 하한 쪽 기각역 확률 Φ(√10−z_{α/2})을 올바르게 더한 식이다.

오답 포인트

양측검정의 검정력은 기각역이 좌우 두 구간으로 나뉘어 있다는 점을 놓치면 틀리기 쉽다. 특히 오른쪽 기각역(큰 쪽) 확률을 구할 때는 $P(Z > c) = 1 - \Phi(c)$ 형태로 여사건을 취해야 하는데, 이를 그대로 $\Phi(c)$로 쓰거나 부호를 반대로 쓰는 실수가 흔하다. 또한 표준화할 때 참값 $\mu = 4$를 기준으로 해야 하는데 귀무가설의 $\mu = 5$를 기준으로 표준화하면 전혀 다른 식이 나온다.

다시 나오면 이렇게 판단

먼저 기각역을 $\bar{X}$ 기준 부등식으로 쓰고, 그 다음 참값(대립가설에서 주어진 μ)으로 표준화하여 두 개의 확률(오른쪽 기각역은 $1 - \Phi(\cdot)$, 왼쪽 기각역은 $\Phi(\cdot)$)을 각각 세운 뒤 더한다. 표준화 상수의 부호(귀무가설 값과 참값의 차이, 여기서는 $5 - 4 = 1$에 $\sqrt{n}$을 곱한 $\sqrt{10}$)가 어느 쪽에 더해지고 어느 쪽에 빼지는지를 부등식 방향으로 확인한다.

암기 포인트

분산 기지 정규모집단, 표본크기 n일 때 양측검정 $H_0 : \mu = \mu_0$의 검정력(참값 μ_1에서)은 $1 - \Phi\left(\sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} + z_{\alpha/2}\right) + \Phi\left(\sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} - z_{\alpha/2}\right)$ 형태로 항상 '오른쪽은 1−Φ, 왼쪽은 Φ'로 짝지어 더하는 구조임을 기억해 둔다.

NoteBase

문 11. 영수는 편의점에서 아메리카노 4개, 카페라떼 3개를 구매하여 상자 하나에 담아 두었다. 그 상자에서 임의로 두 개의 커피를 꺼낼 때, 아메리카노가 한 개 이하일 확률은?

① $\frac{5}{7}$
② $\frac{6}{7}$
③ $\frac{5}{8}$
④ $\frac{3}{4}$

정답 1번

출제 포인트

서로 다른 두 종류의 물건이 섞인 상자에서 두 개를 비복원으로 꺼낼 때 특정 개수 이하가 나올 확률을 초기하분포(조합)로 계산하는 문제이다. "한 개 이하"라는 조건을 0개인 경우와 1개인 경우로 나누어 더해야 한다는 점이 핵심이다.

조건 정리

상자 안 커피는 아메리카노 4개, 카페라떼 3개로 총 7개이다. 이 중 임의로 2개를 꺼낸다(비복원, 순서 무관). "아메리카노가 한 개 이하"인 사건은 아메리카노 0개(카페라떼 2개)이거나 아메리카노 1개(카페라떼 1개)인 경우이다.

사용할 공식

전체 경우의 수는 $\binom{7}{2}$이다. 아메리카노 k개, 카페라떼 $2-k$개를 뽑는 경우의 수는 $\binom{4}{k}\binom{3}{2-k}$이며, 구하는 확률은 $P(X \le 1) = \frac{\binom{4}{0}\binom{3}{2} + \binom{4}{1}\binom{3}{1}}{\binom{7}{2}}$ 이다(여기서 X는 뽑힌 아메리카노의 개수).

계산 과정

전체 경우의 수는 $\binom{7}{2} = 21$이다. 아메리카노 0개(카페라떼 2개)인 경우의 수는 $\binom{4}{0}\binom{3}{2} = 1 \times 3 = 3$이다. 아메리카노 1개(카페라떼 1개)인 경우의 수는 $\binom{4}{1}\binom{3}{1} = 4 \times 3 = 12$이다. 따라서 아메리카노가 한 개 이하인 경우의 수는 $3 + 12 = 15$이고, 확률은 $P(X \le 1) = \frac{15}{21} = \frac{5}{7}$ 이다.

선지 판단

① 5/7: 계산 결과와 일치한다. 정답이다.
② 6/7: 아메리카노가 2개인 경우(여사건)까지 잘못 포함하거나 분모·분자를 잘못 계산한 값으로, 오답이다.
③ 5/8: 전체 경우의 수를 $\binom{7}{2} = 21$ 대신 다른 값(예: 8을 분모로 사용)으로 잘못 잡은 값으로, 오답이다.
④ 3/4: 아메리카노가 2개일 확률의 여사건을 잘못 구하거나 계산 과정에서 항을 누락한 값으로, 오답이다.

NoteBase

오답 포인트

"한 개 이하"를 아메리카노가 정확히 1개인 경우로만 계산하면 12/21=4/7이 나와 어느 선지와도 맞지 않는데, 이는 0개인 경우를 빠뜨린 실수이다. 반대로 여사건(아메리카노 2개)을 구해 1에서 빼는 방법을 쓸 때 $\binom{4}{2}\binom{3}{0} = 6$을 21로 나누어 $1 - 6/21 = 15/21 = 5/7$로 검산하면 같은 답이 나오므로, 두 풀이가 서로 다른 값을 주면 조합 계산(특히 $\binom{4}{2} = 6$, $\binom{3}{2} = 3$)에서 실수한 것이다.

다시 나오면 이렇게 판단

"몇 개 이하" 또는 "몇 개 이상" 조건이 나오면 곧바로 여러 경우를 합산해야 함을 인지하고, 직접 더하는 방법과 여사건을 이용하는 방법 두 가지로 각각 계산해 서로 값이 일치하는지 검산하면 실수를 줄일 수 있다.

암기 포인트

비복원으로 두 종류가 섞인 대상에서 몇 개를 꺼내는 확률은 $\binom{n}{k}$를 이용한 초기하분포 형태로 계산하며, 전체 경우의 수는 항상 전체 개수꺼내는 개수임을 기억해 둔다.

NoteBase

문 12. 당뇨병, 고혈압, 고지혈증에 대한 결합확률분포가 다음과 같다. 고지혈증이 있는 환자 중에 고혈압이 없을 조건부 확률 P(고혈압 무 | 고지혈증 유)는?

	당뇨병 유		당뇨병 무	
	고지혈증 유	고지혈증 무	고지혈증 유	고지혈증 무
고혈압 유	0.10	0.01	0.10	0.02
고혈압 무	0.02	0.05	0.10	0.60

① $\frac{5}{16}$
② $\frac{3}{8}$
③ $\frac{7}{16}$
④ $\frac{11}{16}$

정답 2번

출제 포인트

세 변수(당뇨병, 고혈압, 고지혈증)의 결합확률표에서 하나의 변수(당뇨병)를 주변화(marginalize)하여 나머지 두 변수 사이의 조건부 확률을 구하는 문제이다. 조건부 확률의 정의 $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$를 표에 정확히 적용할 수 있는지를 확인한다.

조건 정리

구해야 할 값은 P(고혈압 무 고지혈증 유)이다. 표의 각 칸은 당뇨병 유/무, 고지혈증 유/무, 고혈압 유/무 세 조건이 모두 결합된 확률이므로, 먼저 당뇨병 유/무를 더해 고지혈증과 고혈압만의 결합확률로 정리한다. 고지혈증 유·고혈압 유: 0.10 + 0.10 = 0.20. 고지혈증 유·고혈압 무: 0.02 + 0.10 = 0.12. 고지혈증 무·고혈압 유: 0.01 + 0.02 = 0.03. 고지혈증 무·고혈압 무: 0.05 + 0.60 = 0.65.

사용할 공식

P(고혈압 무 고지혈증 유)=(P(고혈압 무, 고지혈증 유))/(P(고지혈증 유)) 분모인 P(고지혈증 유)는 고지혈증 유에 해당하는 모든 칸(고혈압 유·무 포함)을 더한 값이다.

계산 과정

P(고지혈증 유)=P(고혈압 유, 고지혈증 유)+P(고혈압 무, 고지혈증 유)=0.20+0.12=0.32이다. 분자는 P(고혈압 무, 고지혈증 유)=0.12이다. 따라서 P(고혈압 무 고지혈증 유)=(0.12)/(0.32)=(12)/(32)=(3)/(8) 이다.

NoteBase

선지 판단

① 5/16(=0.3125)는 0.12/0.32 계산값(0.375)과 다르므로 오답이다. 분모를 잘못 잡거나(예: 전체 표의 합 중 일부만 사용) 계산 과정에서 오차가 생긴 값이다.
② 3/8(=0.375): 0.12/0.32=0.375이므로 정확히 일치한다. 정답이다.
③ 7/16(=0.4375)는 분자·분모 값을 다르게 조합했을 때 나오는 값으로, 올바른 계산과 맞지 않아 오답이다.
④ 11/16(=0.6875)는 고지혈증 유 조건에서 고혈압 유의 조건부 확률(0.20/0.32 = 0.625)과도 다르고, 요구된 조건부 확률과 일치하지 않아 오답이다.

오답 포인트

당뇨병 변수를 더해 없애지 않고 표의 한 칸(예: 0.02 또는 0.10)만 분자로 쓰면 당뇨병 조건이 섞여 들어가 오답이 나온다. 또한 분모를 전체 확률 1로 착각하거나, 고혈압 유·무를 바꿔 계산하면 답이 뒤바뀌므로 조건부 확률에서 '어느 것이 조건(분모)이고 어느 것이 사건(분자)인지'를 표기 순서 그대로 정확히 대응시켜야 한다.

다시 나오면 이렇게 판단

세 변수 결합확률표가 나오면 먼저 불필요한 변수(여기서는 당뇨병)를 행/열 방향으로 더해 두 변수만의 결합확률표로 축소한다. 그 다음 조건부 확률은 항상 '조건에 해당하는 칸들의 합'을 분모로, '조건과 사건이 동시에 만족되는 칸의 합'을 분자로 놓고 나눈다.

암기 포인트

조건부 확률 $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$에서 분모는 조건이 되는 사건의 주변확률(marginal probability) 전체 합이라는 점을 기억한다. 다변량 결합확률표는 구하려는 두 변수만 남기고 나머지는 모두 더해 소거한 뒤 계산을 시작하면 실수를 줄일 수 있다.

NoteBase

문 13. 두 지역 A, B에서 새로운 복지 제도에 대한 찬성 여부를 조사한 결과표 일부가 다음과 같다. 지역과 찬성 여부가 독립이라고 가정할 때 (가)의 값은?

찬성 여부 \ 지역	지역 A	지역 B
찬성	25	35
반대	15	(가)

① 10

② 21

③ 25

④ 30

정답 2번

출제 포인트

지역(A, B)과 찬성 여부가 서로 독립이라는 조건을 교차표(분할표)에 적용해 빈칸 (가)의 값을 구하는 문제이다. 두 범주형 변수가 독립이면 한 변수의 조건부 분포가 다른 변수의 값에 관계없이 동일해야 한다는 점을 이용한다.

조건 정리

지역 A: 찬성 25, 반대 15 (합계 40). 지역 B: 찬성 35, 반대 (가). 지역과 찬성 여부가 독립이라고 가정한다.

사용할 공식

두 범주형 변수 X(지역), Y(찬성 여부)가 독립이면 조건부 비율이 지역에 관계없이 같아야 한다: 찬성_A반대_A = 찬성_B반대_B 이 성립해야 한다.

계산 과정

지역 A의 찬성 대 반대 비율은 25 : 15 = 5 : 3이다. 독립 가정에 따라 지역 B에서도 같은 비율이 유지되어야 하므로 35:(가)=5:3이 성립한다. 따라서 (가)=35×(3)/(5)=21이다.

NoteBase

선지 판단

① 10: 비율을 25 : 15가 아닌 다른 방식(예: 반대 인원끼리의 단순 비교)으로 잘못 계산했을 때 나오는 값으로 독립 조건을 만족하지 못한다.
② 21: 35:(가)=5:3을 풀면 (가)=21이 되어 35 : 21 = 5 : 3 = 25 : 15로 지역과 찬성 여부가 독립이 되는 값이다. 정답이다.
③ 25: 지역 A의 찬성자 수 25를 그대로 옮겨 쓴 값으로 비율 조건을 만족하지 않는다.
④ 30: 반대 인원의 단순 평균이나 다른 임의 계산으로 나올 수 있는 값이나 35 : 30은 5 : 3이 아니므로 독립 조건에 어긋난다.

오답 포인트

이 문제의 핵심은 '독립'을 '찬성/반대 인원수가 같다'거나 '합계가 같다'는 식으로 오해하지 않는 것이다. 독립은 어디까지나 비율(조건부 분포)이 같다는 뜻이므로, 반드시 같은 지역 내에서의 찬성 대 반대 비를 구한 뒤 다른 지역에도 그대로 적용해야 한다. 또한 기대도수 공식 행합×열합전체합을 쓰더라도 결국 같은 비율 관계로 귀결되므로, 두 지역의 찬성:반대 비가 동일하다는 사실만 놓치지 않으면 어떤 방식으로 풀어도 21이 나온다.

다시 나오면 이렇게 판단

분할표에서 '독립'이라는 조건과 함께 빈칸을 채우라는 문제가 나오면, 먼저 완전히 채워진 한쪽 행(또는 열)에서 비율을 구하고, 그 비율을 다른 행(또는 열)에 그대로 적용해 빈칸을 구한다. 계산이 복잡해 보이면 기대도수 공식 행합×열합전체합으로 교차검산한다.

암기 포인트

두 범주형 변수가 독립이면 분할표의 각 행(또는 열)에서 다른 변수에 대한 비율이 모두 동일하다: 찬성_A반대_A=찬성_B반대_B.

NoteBase

문 14. 어느 지역에서 남자와 여자 각각 150명을 임의추출하여 특정 도시개발계획에 대한 찬성 여부를 조사한 결과가 다음과 같다.

성별 \ 찬성 여부	찬성	반대	합계
남자	80	70	150
여자	70	80	150

다음 가설을 유의수준 α에서 검정하고자 할 때, 검정 방법과 임계값(critical value)을 옳게 짝지은 것은? (단, z_α와 $\chi^2_\alpha(r)$는 각각 표준정규분포와 자유도가 r인 카이제곱분포의 제$100\times(1-\alpha)$백분위수이다)

H_0: 남자의 찬성률 = 여자의 찬성률
H_1: 남자의 찬성률 > 여자의 찬성률

① (Z 검정, $z_{\alpha/2}$)
② (Z 검정, z_α)
③ (χ^2 검정, $\chi^2_{\alpha/2}(1)$)
④ (χ^2 검정, $\chi^2_\alpha(2)$)

정답 2번

출제 포인트

두 모비율(남자의 찬성률, 여자의 찬성률)의 차이를 검정하는 상황에서, 대립가설이 방향성을 가진 단측검정(H1: 남자의 찬성률 > 여자의 찬성률)일 때 어떤 검정 방법과 임곗값을 사용해야 하는지를 묻는다. 두 표본 비율 차이 검정은 표본크기가 충분히 크면 정규근사를 이용한 Z 검정으로 수행하며, 검정통계량의 방향(대소)까지 판단해야 하므로 양측검정용 통계량인 χ^2 검정과는 구분해야 한다는 점이 핵심이다.

조건 정리

남자 표본 150명 중 찬성 80명, 반대 70명. 여자 표본 150명 중 찬성 70명, 반대 80명. 두 집단은 서로 독립인 두 모집단에서 각각 임의추출한 표본이며, 각 표본크기(150명)가 충분히 커서 표본비율의 분포를 정규분포로 근사할 수 있다. 가설은 H0: 남자의 찬성률 = 여자의 찬성률, H1: 남자의 찬성률 > 여자의 찬성률로, 방향이 있는 단측(우측) 검정이다.

NoteBase

사용할 공식

두 모비율 차이에 대한 검정통계량은 합동표본비율 $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$을 이용하여 $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$로 정의되며, 표본크기가 충분히 크면 Z는 근사적으로 표준정규분포를 따른다. H1이 한쪽 방향(>)만을 주장하는 단측검정이므로 기각역은 분포의 한쪽 꼬리에만 설정되고, 임곗값은 유의수준 α 전체를 한쪽에 몰아준 z_α가 된다. 반면 χ^2 적합도·독립성 검정 통계량 $\sum \frac{(O-E)^2}{E}$은 편차를 제곱한 형태라서 방향 정보가 사라지므로, '어느 쪽이 더 큰가'를 구분해야 하는 단측 대립가설에는 사용할 수 없다.

계산 과정

검정 방법 판단: 두 독립표본 비율 비교이고 각 표본크기 150명으로 정규근사 조건(np, n(1-p) 모두 충분히 큼)을 만족하므로 Z 검정이 적합하다. 임곗값 판단: H1이 '>'로 한쪽 방향만 제시하는 단측검정이므로, 유의수준 α를 분포의 한쪽 꼬리에 전부 배정한 임곗값 z_α를 사용한다. 만약 H1이 '≠'였다면 양쪽 꼬리에 $\alpha/2$씩 배정하는 $z_{\alpha/2}$를 사용했을 것이다. χ^2 검정은 두 비율이 '같은가 다른가'만 판단하는 양측적 성격의 검정이라 방향이 있는 이 문제의 대립가설과 맞지 않는다.

선지 판단

① (Z 검정, $z_{\alpha/2}$): X. 검정 방법은 Z 검정으로 맞으나, H1이 단측(>)이므로 양측 임곗값 $z_{\alpha/2}$가 아니라 단측 임곗값 z_α를 써야 한다.
② (Z 검정, z_α): O. 두 모비율 차이 검정이며 표본크기가 커서 정규근사가 가능하므로 Z 검정이 맞고, H1이 한쪽 방향(>)만 주장하는 단측검정이므로 임곗값은 z_α가 맞다.
③ (χ^2 검정, $\chi^2_{\alpha/2}(1)$): X. χ^2 검정 자체가 방향을 구분하지 못해 이 문제의 단측 대립가설에 부적합하고, 애초에 χ^2 분포는 좌우 대칭이 아니어서 $\alpha/2$ 형태의 임곗값 개념도 통상적으로 쓰이지 않는다.
④ (χ^2 검정, $\chi^2_\alpha(2)$): X. χ^2 검정 자체가 부적합하며, 2×2 분할표의 독립성 검정 자유도는 $(2-1)(2-1)=1$이지 2가 아니다.

오답 포인트

가장 흔한 실수는 대립가설의 부등호 방향(>)을 놓치고 무조건 $z_{\alpha/2}$를 고르는 것이다. 임곗값이 $\alpha/2$인지 α인지는 대립가설이 '≠'(양측)인지 '>' 또는 '<'(단측)인지에 따라 결정된다는 점을 반드시 확인해야 한다. 또한 2×2 분할표라고 해서 무조건 χ^2 검정을 떠올리기 쉬운데, χ^2 검정은 방향성이 없는 검정이므로 '어느 집단의 비율이 더 큰가'를 묻는 단측 대립가설에는 원칙적으로 쓸 수 없다는 점도 함께 기억해야 한다.

다시 나오면 이렇게 판단

먼저 대립가설(H1)의 부등호를 확인한다. '≠'이면 양측검정이므로 $z_{\alpha/2}$(또는 $\chi^2_{\alpha/2}$), '>' 또는 '<'이면 단측검정이므로 z_α를 쓴다. 다음으로 검정 대상이 '두 비율의 차이'이고 방향까지 판단해야 하면 Z 검정, 여러 범주의 '적합도·독립성(같다/다르다)'만 판단하면 χ^2 검정을 선택한다는 원칙을 순서대로 적용한다.

NoteBase

암기 포인트

단측검정(>, <)은 임곗값에 α 전체, 양측검정($\neq$)은 임곗값에 $\alpha/2$를 쓴다. 두 비율 차이의 방향을 판단하는 검정은 Z 검정, χ^2 검정은 방향이 없는 동질성·독립성 검정에만 쓴다.

NoteBase

문 15. 여러 종류의 직물에 대해 인화 시간(단위: 초)을 측정한 자료를 이용하여 직물에 따라 인화 시간의 모평균이 다른지를 검정하기 위한 일원배치 분산분석의 결과가 다음과 같을 때, 이에 대한 설명으로 옳지 않은 것은? (단, $F_\alpha(k_1, k_2)$는 자유도가 (k_1, k_2)인 F-분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $F_{0.01}(2, 35) = 5.27$, $F_{0.05}(2, 35) = 3.27$, $F_{0.1}(2, 35) = 2.46$이다)

요인	제곱합	자유도	평균제곱	F-값
직물	72	2		
오차	315	35		
전체	387	37		

① 이 실험에 사용된 직물의 종류는 3개이다.

② 직물에 대한 F-값은 4이다.

③ 유의수준 10%에서 직물에 따라 인화 시간에 차이가 없다.

④ 이 가설검정에 대한 p-값은 0.01보다 크고 0.05보다 작다.

정답 3번

출제 포인트

일원배치 분산분석표의 제곱합과 자유도만 주어졌을 때 평균제곱과 F-값을 직접 계산하고, 주어진 F-분포 백분위수와 비교하여 유의성 여부와 p-값의 범위를 판단할 수 있는지를 묻는다.

조건 정리

요인(직물)의 제곱합은 72, 자유도는 2이다. 오차의 제곱합은 315, 자유도는 35이다. 전체 제곱합은 387, 전체 자유도는 37이다. $F_{0.1}(2, 35) = 2.46$, $F_{0.05}(2, 35) = 3.27$, $F_{0.01}(2, 35) = 5.27$이 주어져 있다.

사용할 공식

요인의 평균제곱은 MS_처리=(SS_처리)/(df_처리), 오차의 평균제곱은 MS_오차=(SS_오차)/(df_오차)이며, 검정통계량은 F=(MS_처리)/(MS_오차)이다. 자유도가 2인 요인의 수준 개수(직물 종류의 수)는 df_처리=k-1에서 k=df_처리+1로 구한다.

NoteBase

계산 과정

직물의 자유도가 2이므로 직물의 종류 수는 $k = 2 + 1 = 3$개이다.
직물의 평균제곱은 MS_직물=(72)/(2)=36이다.
오차의 평균제곱은 MS_오차=(315)/(35)=9이다.
검정통계량은 $F = \frac{36}{9} = 4$이다.
이제 계산된 $F = 4$를 주어진 백분위수와 비교한다.
$F_{0.1}(2, 35) = 2.46 < F_{0.05}(2, 35) = 3.27 < F = 4 < F_{0.01}(2, 35) = 5.27$의 순서가 성립한다.
$F = 4$는 유의수준 5%의 기각값 3.27보다 크므로 유의수준 5%(그리고 이보다 큰 유의수준 10%)에서도 귀무가설(모든 직물의 인화 시간 모평균이 같다)을 기각한다. 즉 유의수준 10%에서 직물에 따라 인화 시간에 차이가 있다고 판정한다.
한편 $F = 4$는 유의수준 1%의 기각값 5.27보다는 작으므로 유의수준 1%에서는 귀무가설을 기각하지 못한다.
유의수준 5%에서는 기각($F = 4 > 3.27$)하고 유의수준 1%에서는 기각하지 못하므로($F = 4 < 5.27$), p-값은 0.01과 0.05 사이에 위치한다.

선지 판단

① 이 실험에 사용된 직물의 종류는 3개이다. → 직물의 자유도가 2이므로 종류 수는 $2 + 1 = 3$개가 맞다. 옳은 설명이다.
② 직물에 대한 F-값은 4이다. → MS_직물=36, MS_오차=9이므로 $F = 36/9 = 4$가 맞다. 옳은 설명이다.
③ 유의수준 10%에서 직물에 따라 인화 시간에 차이가 없다. → $F = 4 > F_{0.1}(2, 35) = 2.46$이므로 유의수준 10%에서 귀무가설을 기각하여 직물에 따라 인화 시간에 차이가 있다고 판정해야 한다. 따라서 '차이가 없다'는 설명은 틀린 표현이며, 올바른 표현은 '유의수준 10%에서 직물에 따라 인화 시간에 차이가 있다'이다. 옳지 않은 설명이므로 이것이 정답이다.
④ 이 가설검정에 대한 p-값은 0.01보다 크고 0.05보다 작다. → $F = 4$는 $F_{0.05}(2, 35) = 3.27$보다 크고 $F_{0.01}(2, 35) = 5.27$보다 작으므로 p-값은 0.01과 0.05 사이에 있다. 옳은 설명이다.

오답 포인트

F-값과 기각값의 대소 비교 방향을 반대로 생각하면 오답에 이른다. 계산된 F-값이 기각값보다 클수록 더 작은 유의수준의 기각값까지 넘어서는 것이므로 p-값은 더 작아진다. 유의수준을 크게 잡을수록(예: 10%) 기각값은 작아지므로 귀무가설을 기각하기 더 쉬워진다는 점을 혼동하지 않도록 한다.

다시 나오면 이렇게 판단

분산분석표에서 제곱합과 자유도만 주어지면 먼저 평균제곱(제곱합/자유도)을 각각 구하고, 그 비로 F-값을 계산한다. 이후 F-값을 주어진 임계값들과 크기 순으로 나열해 어느 구간에 있는지 확인하면 유의수준별 기각 여부와 p-값의 범위를 동시에 판단할 수 있다. 요인의 자유도가 $k - 1$이라는 관계로부터 수준(집단) 개수도 바로 구할 수 있다.

NoteBase

암기 포인트

F=(MS_처리)/(MS_오차)=(SS_처리/df_처리)/(SS_오차/df_오차); 요인의 자유도 df_처리=k-1(k=수준 수); F-값이 유의수준 α의 기각값보다 크면 그 유의수준에서 귀무가설을 기각하고, F-값이 두 기각값 사이에 있으면 p-값도 그 두 유의수준 사이에 위치한다.

NoteBase

문 16. 다음 상황에 대한 검정 방법으로 옳지 않은 것은? (단, 각 검정에 대해 정규성과 등분산성을 만족하고 검정에 필요한 분산은 알려지지 않은 것으로 한다)

새로 개발된 당뇨치료제의 효과를 알아보기 위한 임상시험에서 자발적으로 지원한 20명의 당뇨 환자 중 무작위로 10명을 선택하여 A 그룹으로 배정하고 나머지를 B 그룹으로 배정한 후, A 그룹에는 개발된 치료제를, B 그룹에는 위약(placebo)을 한 달간 복용하게 하였다. 각 그룹에 대한 치료 효과를 비교하기 위해 복용 전의 혈당량과 복용 후의 혈당량을 측정하였다.

① A 그룹과 B 그룹의 복용 전 평균 혈당량의 비교는 독립 t-검정으로 실시한다.

② A 그룹 내에서 각 환자의 복용 전과 후의 혈당량 비교는 대응(쌍체) t-검정으로 실시한다.

③ 각 환자의 복용 전과 후의 혈당량 차이에 대한 A 그룹과 B 그룹 간의 평균 비교는 독립 t-검정으로 실시한다.

④ 복용 후 혈당량에 대한 A 그룹과 B 그룹 간의 평균 비교는 대응(쌍체) t-검정으로 실시한다.

정답 4번

출제 포인트

같은 대상을 반복 측정한 값(대응·쌍체 자료)인지, 서로 다른 두 독립 집단에서 얻은 값인지를 구분해 대응(쌍체) t-검정과 독립 t-검정 중 무엇을 써야 하는지 판단하는 문제이다. A 그룹과 B 그룹은 20명의 환자를 무작위로 나눈 서로 다른 사람들이므로 두 그룹 사이의 비교는 항상 독립 표본이고, 같은 환자를 복용 전·후로 두 번 잰 값끼리의 비교만 대응 표본이 된다.

핵심 개념

대응(쌍체) t-검정은 동일한 개체를 두 번(또는 짝지어) 측정해 그 차이를 분석할 때 쓴다. 예를 들어 같은 환자의 복용 전 혈당량과 복용 후 혈당량처럼 측정값끼리 자연스러운 짝이 존재해야 한다. 독립 t-검정은 서로 관련 없는 두 집단의 평균을 비교할 때 쓴다. A 그룹 환자와 B 그룹 환자는 무작위 배정으로 나뉜 별개의 사람들이므로 한 그룹의 값과 다른 그룹의 값 사이에는 짝지어짐이 존재하지 않는다. 따라서 A와 B 사이의 어떤 비교든(복용 전 평균끼리, 복용 후 평균끼리, 혹은 전후 차이의 평균끼리) 모두 독립 t-검정의 대상이고, 대응 t-검정은 오직 같은 그룹 내에서 같은 환자의 전·후 값을 비교할 때만 성립한다.

NoteBase

선지별 해설

① A 그룹과 B 그룹의 복용 전 평균 혈당량 비교 → A와 B는 서로 다른 환자들로 구성된 독립 집단이므로 두 집단의 평균을 비교하는 것은 독립 t-검정이 맞다. (O)
② A 그룹 내에서 각 환자의 복용 전과 후의 혈당량 비교 → 같은 환자를 두 번(전, 후) 측정한 값이므로 자연스러운 짝이 존재한다. 대응(쌍체) t-검정이 맞다. (O)
③ 각 환자의 복용 전과 후의 혈당량 차이(전후 차)를 A 그룹과 B 그룹 간에 비교 → 비교 대상이 '전후 차이'라는 값으로 바뀌었을 뿐, 그 값을 낸 환자들은 여전히 A 그룹과 B 그룹이라는 서로 다른 두 집단에 속한다. 따라서 이 두 집단의 평균(전후 차의 평균)을 비교하는 것은 독립 t-검정이 맞다. (O)
④ 복용 후 혈당량에 대한 A 그룹과 B 그룹 간의 평균 비교 → A 그룹과 B 그룹은 서로 다른 환자들로 구성된 독립 집단이므로 짝지어진 관계가 없다. 그런데 이 선지는 이를 대응(쌍체) t-검정으로 실시한다고 하였으므로 틀렸다. 올바른 방법은 독립 t-검정이다. 틀린 표현 → 올바른 표현: '대응(쌍체) t-검정으로 실시한다' → '독립 t-검정으로 실시한다'. (X)

오답 포인트

이 문제에서 흔히 하는 착각은 '같은 실험, 같은 그룹 이름(A, B)이 반복해서 나오니 다 대응 검정이겠지'라고 생각하는 것이다. 하지만 대응 여부는 그룹 이름이 아니라 '누구를 비교하는가'로 결정된다. A 그룹과 B 그룹은 애초에 무작위로 나뉜 서로 다른 사람들이므로, 무엇을 측정값으로 삼든(전 혈당량이든 후 혈당량이든 전후 차이든) A와 B 사이의 비교는 항상 독립 t-검정이다. 반대로 같은 환자 내에서 전과 후를 비교할 때만 대응 t-검정이 된다.

다시 나오면 이렇게 판단

비교하는 두 값이 '같은 개체에서 나온 짝'인지 '서로 다른 개체에서 나온 값'인지를 먼저 따진다. 같은 환자의 전·후 값 비교 → 대응(쌍체) t-검정. 서로 다른 그룹(무작위로 나뉜 별개의 사람들) 간의 평균 비교 → 독립 t-검정. 비교 대상이 원래 측정값이든 전후 차이라는 파생값이든, 그 값이 속한 단위(같은 사람인지 다른 그룹인지)를 기준으로 판단한다.

암기 포인트

같은 사람, 두 번 측정 = 대응(쌍체) t-검정. 다른 사람, 다른 그룹 = 독립 t-검정. A 그룹과 B 그룹 사이의 비교는 무엇을 비교하든 항상 독립 t-검정이라는 점을 기억해 둔다.

NoteBase

문 17. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ i = 1, 2, \cdots, n$에서 β_0와 β_1의 최소제곱추정량을 각각 $\hat{\beta}_0$와 $\hat{\beta}_1$이라 하고 잔차 e_i를 $Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$이라 할 때, 이에 대한 설명으로 옳은 것만을 모두 고른 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

ㄱ. e_i와 x_i의 곱의 합은 0이다.
ㄴ. $\hat{\beta}_1$의 부호는 설명변수와 반응변수 간의 표본상관계수의 부호와 같다.
ㄷ. 결정계수의 부호와 $\hat{\beta}_1$의 부호는 같다.
ㄹ. 결정계수의 부호는 설명변수와 반응변수 간의 표본상관계수의 부호와 같다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄹ
④ ㄷ, ㄹ

정답 1번

출제 포인트

단순선형회귀모형에서 최소제곱추정량 $\hat{\beta}_0, \hat{\beta}_1$과 잔차 e_i가 만족하는 정규방정식의 성질, 그리고 $\hat{\beta}_1$·표본상관계수·결정계수 세 값 사이의 부호·범위 관계를 정확히 아는지를 묻는 문제다. 특히 결정계수는 항상 0 이상 1 이하의 값이라 애초에 '부호'라는 개념 자체가 성립하지 않는다는 점이 핵심 함정이다.

조건 정리

모형은 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$이고 서로 독립이다. $\hat{\beta}_0, \hat{\beta}_1$은 최소제곱추정량이고 잔차는 $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$로 정의된다.

사용할 공식

최소제곱법의 정규방정식: $\sum e_i = 0, \sum e_i x_i = 0$. 기울기 추정량과 상관계수의 관계: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, 표본상관계수 $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$ (단 $S_{xy} = \sum (x_i - \bar{x})(Y_i - \bar{Y}), S_{xx} = \sum (x_i - \bar{x})^2 > 0, S_{yy} = \sum (Y_i - \bar{Y})^2 > 0$). 결정계수는 단순선형회귀에서 $R^2 = r^2$이다.

NoteBase

계산 과정

ㄱ의 검토: 최소제곱추정량은 잔차제곱합 $\sum e_i^2$을 β_0, β_1에 대해 미분해 0으로 놓아 얻어지는데, 이 두 정규방정식이 바로 $\sum e_i = 0$과 $\sum e_i x_i = 0$이다. 따라서 잔차와 x_i의 곱의 합은 항상 0이다. ㄱ은 옳다.
ㄴ의 검토: $\hat{\beta}_1 = S_{xy}/S_{xx}$이고 $r = S_{xy}/(\sqrt{S_{xx}}\sqrt{S_{yy}})$이므로 두 식 모두 분자가 S_{xy}이고 분모는 항상 양수($S_{xx} > 0, \sqrt{S_{xx}}\sqrt{S_{yy}} > 0$)이다. 따라서 $\hat{\beta}_1$의 부호는 오직 S_{xy}의 부호에 의해 결정되며, 이는 r의 부호와 정확히 같다. ㄴ은 옳다.
ㄷ의 검토: 단순선형회귀에서 결정계수는 $R^2 = r^2 \geq 0$이므로 결정계수는 음수가 될 수 없다. 즉 결정계수는 '부호'라는 개념 자체가 없는 항상 0 이상인 값이다. 그런데 $\hat{\beta}_1$은 S_{xy}의 부호에 따라 음수가 될 수 있으므로, $\hat{\beta}_1$이 음수인 경우 결정계수(항상 양수 또는 0)와 부호가 같다고 말할 수 없다. ㄷ은 틀렸다.
ㄹ의 검토: 같은 이유로 결정계수 $R^2 = r^2$은 항상 0 이상이지만 표본상관계수 r은 음수일 수 있으므로, $r < 0$인 경우 결정계수의 부호(항상 비음수)와 r의 부호(음수)가 같을 수 없다. ㄹ은 틀렸다.

선지 판단

① ㄱ, ㄴ → O: 두 진술 모두 최소제곱법의 정규방정식과 β_1=r·(양수 비율) 관계로부터 항상 성립하는 옳은 설명이다.
② ㄱ, ㄷ → X: ㄱ은 옳지만 ㄷ은 결정계수가 항상 비음수라는 점에서 틀렸다.
③ ㄴ, ㄹ → X: ㄴ은 옳지만 ㄹ은 결정계수가 항상 비음수라는 점에서 틀렸다.
④ ㄷ, ㄹ → X: ㄷ, ㄹ 모두 결정계수에 '부호'를 대응시키는 것 자체가 틀린 설명이다.

오답 포인트

결정계수(R^2)를 회귀계수나 상관계수와 같은 '부호가 있는 값'으로 착각하면 ㄷ, ㄹ을 옳다고 고르게 된다.
결정계수는 정의상 제곱값이므로 항상 0 이상 1 이하이며, 음의 방향을 가질 수 없다는 점을 놓치지 않아야 한다.

다시 나오면 이렇게 판단

선택지에 '결정계수의 부호'라는 표현이 나오면 먼저 결정계수가 r^2 형태로 항상 비음수라는 사실부터 확인한다.
그 다음 $\hat{\beta}_1$과 표본상관계수는 분모가 항상 양수인 비례식이므로 서로 부호가 같다는 점, 그리고
최소제곱추정량은 항상 $\sum e_i = 0, \sum e_i x_i = 0$이라는 정규방정식을 만족한다는 점을 기본 성질로 떠올린다.

암기 포인트

정규방정식: $\sum e_i = 0, \sum e_i x_i = 0$. 부호 관계: $\text{sign}(\hat{\beta}_1) = \text{sign}(r)$. 범위 성질: $R^2 = r^2 \geq 0$이므로 결정계수는 부호를 갖지 않는다.

NoteBase

문 18. 다음 중 선형회귀모형을 적용할 수 있는 모형만을 모두 고른 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

ㄱ. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
ㄴ. $Y_i = e^{\beta_0} \times x_i^{\beta_1} \times e^{\varepsilon_i}$
ㄷ. $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

① ㄱ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

선형회귀모형의 '선형성'은 설명변수 x에 대한 선형성이 아니라 모수(파라미터) β_0, β_1에 대한 선형성을 의미한다는 점을 묻는 문제이다. 식이 겉보기에 지수함수나 이차항을 포함하더라도, 변수를 적절히 변환했을 때 모수에 대해 선형이고 오차항이 가법적(더하기 형태)으로 들어오면 선형회귀모형으로 취급할 수 있다.

조건 정리

세 모형 ㄱ, ㄴ, ㄷ이 선형회귀모형(모수에 대해 선형이고 오차항이 가법적)으로 적용 가능한지를 각각 판단해야 한다. 오차항 ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이라는 표준적인 선형회귀 가정을 만족한다.

사용할 공식

일반적인 선형회귀모형의 형태는 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$이며, 모수 β_0, β_1에 대해 1차식이어야 한다. 지수형 모형 $Y_i = e^{\beta_0} \times x_i^{\beta_1} \times e^{\varepsilon_i}$은 양변에 자연로그를 취하면 $\ln Y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$가 되어, $\ln Y_i$를 종속변수로, $\ln x_i$를 설명변수로 치환하면 모수에 대해 선형인 식으로 바뀐다.

NoteBase

계산 과정

ㄱ. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$: 모수 β_0, β_1에 대해 1차식이고 오차항이 더해지는 형태이므로 표준적인 선형회귀모형 그 자체이다.

ㄴ. $Y_i = e^{\beta_0} \times x_i^{\beta_1} \times e^{\varepsilon_i}$: 양변에 자연로그를 취하면 $\ln Y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$가 된다. 새로운 변수 $Y_i' = \ln Y_i$, $x_i' = \ln x_i$로 두면 $Y_i' = \beta_0 + \beta_1 x_i' + \varepsilon_i$ 꼴이 되어 모수 β_0, β_1에 대해 선형이고 오차항도 가법적으로 들어간다. 따라서 로그변환을 통해 선형회귀모형으로 적용할 수 있다.

ㄷ. $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$: 설명변수 자체는 x_i^2으로 비선형이지만, 새로운 변수 $x_i' = x_i^2$로 치환하면 $Y_i = \beta_0 + \beta_1 x_i' + \varepsilon_i$ 꼴이 되어 모수 β_0, β_1에 대해 여전히 1차식이다. 선형회귀모형에서 말하는 선형성은 설명변수가 아닌 모수에 대한 선형성이므로, 이 모형도 선형회귀모형으로 적용할 수 있다.

세 모형 모두 모수에 대해 선형인 형태로 변환 가능하므로 전부 적용 가능하다.

선지 판단

① ㄱ → X. ㄱ뿐 아니라 ㄴ, ㄷ도 모수에 대해 선형이어서 선형회귀모형에 해당하므로 ㄱ만 고르는 것은 불완전하다.
② ㄱ, ㄷ → X. ㄴ을 제외했으나, ㄴ도 로그변환을 거치면 모수에 대해 선형인 식이 되므로 제외할 근거가 없다.
③ ㄴ, ㄷ → X. ㄱ을 제외했으나, ㄱ은 애초에 표준적인 선형회귀모형이므로 당연히 포함되어야 한다.
④ ㄱ, ㄴ, ㄷ → O. 세 모형 모두 (필요시 변수변환을 거쳐) 모수 β_0, β_1에 대해 선형이고 오차항이 가법적인 형태로 표현되므로 전부 선형회귀모형으로 적용할 수 있다.

오답 포인트

가장 흔한 오답은 ㄴ을 '지수함수라서 비선형모형'이라고 단정해 제외하는 것이다(②, ③ 선택). 그러나 선형회귀모형의 '선형'은 설명변수에 대한 선형이 아니라 모수(회귀계수)에 대한 선형을 뜻하므로, 로그변환처럼 함수변환을 거쳐 모수에 대해 선형인 식으로 바뀌면 선형회귀모형으로 분석할 수 있다. 같은 이유로 ㄷ의 x_i^2 항도 변수 하나로 치환하면 선형이 되므로 비선형모형으로 오인하면 안 된다.

다시 나오면 이렇게 판단

모형이 나오면 가장 먼저 회귀계수(모수)에 대해 1차식인지부터 확인한다. 설명변수 자리에 x^2, $\ln x$, $\sqrt{x}$ 등 변수변환이 들어가 있어도 계수 β_0, β_1이 서로 곱해지거나 지수에 있지 않고 단순히 더해지는 형태라면 선형회귀모형으로 볼 수 있다. 반대로 β_1이 지수나 분모에 있어 로그변환 등으로도 선형화되지 않는 경우에만 비선형회귀모형으로 판단한다.

NoteBase

암기 포인트

선형회귀모형의 '선형'은 변수(x)가 아니라 모수(β)에 대한 선형성이다. 지수형 모형은 로그변환, 다항형 모형(제곱항 등)은 변수치환을 통해 모수에 대해 선형인 형태로 바꿀 수 있으면 선형회귀모형으로 적용 가능하다.

NoteBase

문 19. 다중선형회귀모형 $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,\ i = 1, 2, \cdots, n$에서 결정계수($R^2$)와 수정결정계수(adjusted R^2)에 대한 설명으로 옳은 것은? (단, $k \geq 1$)

① 동일한 설명변수를 사용할 때 수정결정계수는 결정계수보다 작거나 같다.

② 설명변수를 추가할수록 결정계수는 작아진다.

③ 설명변수를 추가할수록 수정결정계수는 커진다.

④ 수정결정계수는 다중공선성의 판단기준이 된다.

정답 1번

출제 포인트

다중선형회귀에서 결정계수 R^2와 수정결정계수(adjusted R^2)가 설명변수의 개수 변화에 따라 각각 어떻게 움직이는지, 그리고 두 지표의 정의상 크기 관계를 정확히 알고 있는지를 묻는다. R^2은 표본에 대한 적합도만 나타내는 반면 수정결정계수는 설명변수 개수(자유도)에 대한 벌점을 반영한다는 점이 핵심이다.

핵심 개념

결정계수는 $R^2 = 1 - \frac{SSE}{SST}$로 정의되며, 설명변수를 추가하면 SSE(잔차제곱합)는 절대 증가하지 않으므로 R^2은 항상 같거나 커진다(감소하지 않는다). 수정결정계수는 $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$로 정의되어 설명변수 개수 k에 대한 벌점(자유도 조정)이 들어간다. 이 식에서 $k \geq 1$이면 $\frac{n-1}{n-k-1} > 1$이므로, 같은 설명변수 조합에 대해 항상 $\bar{R}^2 \leq R^2$이 성립한다(등호는 $R^2 = 1$일 때만). 또한 설명변수를 추가할 때 R^2의 증가폭이 벌점 증가분을 상쇄하지 못하면 수정결정계수는 오히려 감소할 수 있으므로, 수정결정계수는 항상 커진다고 단정할 수 없다. 다중공선성 여부는 분산팽창인자(VIF)나 상관행렬 등으로 판단하는 것이지 수정결정계수로 판단하는 것이 아니다.

선지별 해설

① 동일한 설명변수를 사용할 때 수정결정계수는 결정계수보다 작거나 같다. → O. 위 식에서 $k \geq 1$이면 벌점항 $\frac{n-1}{n-k-1}$이 1보다 크므로 $\bar{R}^2 \leq R^2$이 항상 성립한다.

② 설명변수를 추가할수록 결정계수는 작아진다. → X. 설명변수를 추가하면 잔차제곱합이 증가하지 않으므로 결정계수는 감소하지 않고 같거나 커진다 → "작아진다"를 "작아지지 않고 같거나 커진다"로 고쳐야 한다.

③ 설명변수를 추가할수록 수정결정계수는 커진다. → X. 수정결정계수는 설명변수 추가에 따른 R^2 증가분과 벌점 증가분의 상대적 크기에 따라 커질 수도, 작아질 수도 있다 → "커진다"를 "커질 수도 작아질 수도 있다"로 고쳐야 한다.

④ 수정결정계수는 다중공선성의 판단기준이 된다. → X. 다중공선성은 분산팽창인자(VIF) 등으로 판단하며 수정결정계수는 모형의 설명력을 자유도 조정하여 나타낸 지표일 뿐이다 → "다중공선성의 판단기준"을 "모형 설명력(적합도)의 비교 기준"으로 고쳐야 한다.

NoteBase

오답 포인트

결정계수는 설명변수를 추가할수록 절대 감소하지 않는다는 성질과, 수정결정계수는 이 성질이 성립하지 않는다는 점을 혼동하기 쉽다. 또한 수정결정계수가 자유도를 조정한 지표라는 이유만으로 다중공선성 진단 도구로 오해하는 경우가 많은데, 다중공선성 진단은 별도의 통계량(VIF 등)으로 이루어진다는 점을 구분해야 한다.

다시 나오면 이렇게 판단

먼저 결정계수는 설명변수 추가에 대해 단조 비감소(같거나 커짐)라는 성질을, 수정결정계수는 그렇지 않다는 성질을 떠올린다. 그다음 두 지표의 크기 관계(수정결정계수 $\leq$ 결정계수, $k \geq 1$일 때)를 공식으로 확인하고, 마지막으로 다중공선성 판단은 수정결정계수가 아니라 별개의 통계량으로 이루어진다는 것을 체크하면 정답을 좁힐 수 있다.

암기 포인트

결정계수는 설명변수를 추가해도 절대 줄지 않는다(같거나 커짐), 수정결정계수는 늘 수도 줄 수도 있다, 그리고 같은 설명변수 조합이면 수정결정계수는 결정계수보다 작거나 같다는 세 가지를 세트로 기억해 둔다. 다중공선성 판단은 수정결정계수가 아니라 분산팽창인자(VIF)라는 것도 함께 기억해 둔다.

NoteBase

문 20. 다중선형회귀모형 $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,\ i = 1, 2, \cdots, 27$에 최소제곱추정법을 적용하여 얻은 분산분석표가 다음과 같을 때, 이에 대한 설명으로 옳은 것만을 모두 고른 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이며, 자유도가 (k_1, k_2)인 F-분포의 제$100 \times (1-\alpha)$ 백분위수를 $F_\alpha(k_1, k_2)$라고 할 때 $F_{0.01}(2, 24) = 5.61$이다)

요인	제곱합	자유도	평균제곱	F-값
회귀	20	2	10	
잔차	15	24	0.625	
계	35	26		

ㄱ. 회귀직선의 결정계수는 $\frac{20}{35}$이다.

ㄴ. 유의수준 1%에서 가설 $H_0 : \beta_1 = 0$ 대 $H_1 : \beta_1 \neq 0$에 대한 검정에서 귀무가설을 기각한다.

ㄷ. 모형의 유의성 검정에서 귀무가설 모형은 $Y_i = \varepsilon_i$이다.

① ㄱ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

다중회귀분석의 분산분석표를 읽고 결정계수를 계산하는 능력, 그리고 분산분석표의 F-검정이 무엇을 검정하는 것인지(개별 회귀계수 검정과 모형 전체의 유의성 검정을 구별)를 정확히 이해하는지를 묻는다. 표에 제시된 제곱합과 자유도만으로 판단 가능한 부분과, 표만으로는 판단할 수 없는 부분을 가려내는 것이 핵심이다.

조건 정리

설명변수가 x_{1i}, x_{2i} 2개인 다중선형회귀모형이고 표본크기는 $n = 27$이다. 분산분석표에서 회귀제곱합(SSR)=20, 자유도=2, 잔차제곱합(SSE)=15, 자유도=24, 총제곱합(SST)=35, 자유도=26이며 잔차의 평균제곱(MSE)=0.625이다. 자유도가 (2,24)인 F분포의 제99백분위수는 $F_{0.01}(2, 24) = 5.61$로 주어져 있다.

NoteBase

사용할 공식

결정계수 $R^2 = \frac{SSR}{SST}$. 모형 전체의 유의성 검정(귀무가설 $H_0 : \beta_1 = \beta_2 = 0$)의 검정통계량은 $F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$이며 이 통계량은 자유도 $(k, n-k-1)$인 F분포를 따른다. 반면 개별 회귀계수 하나(β_1)에 대한 검정 $H_0 : \beta_1 = 0$은 그 계수의 추정량과 표준오차를 이용한 t-검정(또는 자유도 $(1, n-k-1)$인 부분 F-검정)으로 수행해야 하며, 분산분석표의 전체 회귀 제곱합만으로는 계산할 수 없다.

계산 과정

ㄱ의 검토: 결정계수는 $R^2 = \frac{SSR}{SST} = \frac{20}{35}$이므로 표에 제시된 제곱합 값을 그대로 대입한 것과 일치한다. 옳은 설명이다.

ㄴ의 검토: 표에 제시된 F값은 회귀제곱합(자유도 2)과 잔차제곱합(자유도 24)으로 계산되는 $F = \frac{MSR}{MSE} = \frac{10}{0.625} = 16$으로, 이는 $H_0 : \beta_1 = \beta_2 = 0$(즉 두 설명변수가 모두 회귀식에 기여하지 않는다는) 모형 전체의 유의성을 검정하는 통계량이다. 그런데 ㄴ은 β_1 하나만에 대한 가설 $H_0 : \beta_1 = 0$을 검정하겠다는 것으로, 이는 별도의 t-검정(또는 자유도가 (1,24)인 부분 F-검정)을 필요로 한다. 이 검정에는 β_1의 추정값과 표준오차, 혹은 x_1만 제외했을 때의 제곱합 감소분이 있어야 하는데 주어진 분산분석표에는 이러한 정보가 없다. 즉 주어진 $F_{0.01}(2, 24) = 5.61$과 표의 F값 16은 β_1 하나만의 유의성 판단에 쓸 수 있는 통계량이 아니므로, 이 표만 가지고 ㄴ의 결론(귀무가설 기각)을 내릴 수 없다. 따라서 ㄴ은 옳지 않다.

ㄷ의 검토: 모형의 유의성 검정은 $H_0 : \beta_1 = \beta_2 = 0$을 검정하는 것이며, 이때 귀무가설 하의 축소모형은 $Y_i = \beta_0 + \varepsilon_i$로 절편 β_0는 남아 있고 두 기울기 계수만 0이 되는 모형이다. ㄷ에서 제시한 $Y_i = \varepsilon_i$는 절편 β_0까지 0이라고 가정한 것이므로 실제 검정하는 귀무가설의 모형과 다르다. 따라서 ㄷ은 옳지 않다.

선지 판단

① ㄱ: O. 결정계수는 SSR/SST=20/35로 표의 값과 정확히 일치한다.
② ㄱ, ㄷ: X. ㄷ이 틀린 설명이므로 성립하지 않는다.
③ ㄴ, ㄷ: X. ㄴ, ㄷ 모두 틀린 설명이므로 성립하지 않는다.
④ ㄱ, ㄴ, ㄷ: X. ㄴ, ㄷ이 틀린 설명이므로 성립하지 않는다.

오답 포인트

가장 흔한 실수는 분산분석표에 제시된 F값(자유도 2,24)을 이용해서 개별 회귀계수 β_1에 대한 검정까지 함께 판단해버리는 것이다. 분산분석표의 F-검정은 여러 설명변수를 묶어서 동시에 유의한지를 보는 전체 모형 검정이고, 개별 계수의 유의성은 별도의 t-검정 통계량(또는 다른 자유도의 부분 F-검정)이 필요하다는 점을 혼동하면 ㄴ을 옳다고 잘못 판단하게 된다. 또한 ㄷ에서 '유의성 검정의 귀무가설 모형'을 절편까지 제거한 $Y_i = \varepsilon_i$로 착각하기 쉬운데, 실제로는 절편 β_0는 남기고 기울기 계수들만 0으로 놓은 $Y_i = \beta_0 + \varepsilon_i$가 귀무가설 모형이다.

NoteBase

다시 나오면 이렇게 판단

분산분석표 문제에서는 먼저 그 표의 F-검정이 정확히 어떤 가설(전체 계수 동시 검정인지, 개별 계수 검정인지)을 검정하는지부터 확인한다. 표에 자유도 (설명변수 개수, n-설명변수개수-1)인 F값만 주어져 있다면 이는 전체 모형의 유의성 검정에만 쓸 수 있고, 개별 계수 하나에 대한 검정에는 그대로 쓸 수 없다는 점을 기억한다. 또한 모형의 유의성 검정에서 귀무가설 하의 모형은 항상 절편은 남기고 나머지 설명변수의 계수들만 0으로 놓은 모형이라는 것을 확인한다.

암기 포인트

결정계수 $R^2 = SSR/SST$. 다중회귀 전체 유의성 검정: $H_0 : \beta_1 = \cdots = \beta_k = 0$, 통계량 $F = MSR/MSE$, 자유도 $(k, n-k-1)$, 귀무가설 하 모형은 $Y_i = \beta_0 + \varepsilon_i$(절편은 유지). 개별 회귀계수 하나의 유의성은 분산분석표의 F-검정이 아니라 별도의 t-검정(또는 자유도 1인 부분 F-검정)으로 판단해야 한다.

NoteBase

문 21. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$, $i = 1, 2, \cdots, n$에서 회귀계수 β_0와 β_1에 대한 최소제곱추정량 $\hat{\beta}_0$과 $\hat{\beta}_1$에 대한 설명으로 옳지 않은 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

① $\hat{\beta}_0$은 β_0의 불편추정량이다.
② $\hat{\beta}_1$은 β_1의 불편추정량이다.
③ $\hat{\beta}_0$의 분산은 $\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \times \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$이다.
④ $\hat{\beta}_1$의 분산은 $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$이다.

정답 3번

출제 포인트

이 문제의 회귀모형은 설명변수를 중심화(centering)한 $Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ 형태이다. 이렇게 중심화된 모형에서 최소제곱추정량 $\hat{\beta}_0, \hat{\beta}_1$의 불편성과 분산 공식이 일반적인(중심화하지 않은) 단순선형회귀모형과 어떻게 달라지는지를 정확히 구분해서 알고 있는지를 묻는다.

조건 정리

모형은 $Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$이고 서로 독립이다. 설명변수가 x_i가 아니라 중심화된 값 $(x_i - \bar{x})$로 들어가 있다는 점이 이 모형의 핵심 조건이다.

사용할 공식

중심화 모형의 최소제곱추정량은 $\hat{\beta}_0 = \bar{Y}, \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$이다. 이때 두 추정량의 분산은 각각 $Var(\hat{\beta}_0) = \frac{\sigma^2}{n}$, $Var(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$이다.

계산 과정

설명변수를 $\bar{x}$만큼 중심화했기 때문에 새 설명변수 $(x_i - \bar{x})$의 평균은 0이 된다. 절편 추정량 $\hat{\beta}_0$은 회귀직선이 $(\bar{x}, \bar{Y})$를 지난다는 사실에서 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \times 0 = \bar{Y}$로 정리된다. $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$이고 각 Y_i의 분산이 σ^2이며 서로 독립이므로 $Var(\hat{\beta}_0) = Var(\bar{Y}) = \frac{\sigma^2}{n}$이 된다. 즉 중심화 모형에서는 절편 추정량의 분산에 $\frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2}$ 항이 붙지 않는다. 그 항은 설명변수를 중심화하지 않은 일반적인 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$에서 $\hat{\beta}_0$의 분산 공식 $Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$에 해당하는 것으로, 이 문제의 중심화 모형에는 적용되지 않는다. 한편 기울기 추정량 $\hat{\beta}_1$은 중심화 여부와 무관하게 항상 $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$로 동일하다.

NoteBase

선지 판단

① O — $\hat{\beta}_0 = \bar{Y}$이고 $E(\bar{Y}) = \beta_0$이므로 $\hat{\beta}_0$은 β_0의 불편추정량이다.
② O — 최소제곱추정량 $\hat{\beta}_1$은 중심화 여부와 관계없이 항상 β_1의 불편추정량이다.
③ X — 중심화 모형에서 $Var(\hat{\beta}_0)$은 $\frac{\sigma^2}{n}$이다. 선지의 식 $\sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$은 중심화하지 않은 일반 모형의 절편 분산 공식이므로 이 문제의 조건에는 맞지 않는 틀린 설명이다. → 올바른 식은 $Var(\hat{\beta}_0) = \frac{\sigma^2}{n}$이다.
④ O — $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$은 중심화 모형과 일반 모형 모두에서 동일하게 성립하는 올바른 식이다.

오답 포인트

가장 흔한 오답 원인은 절편과 기울기의 분산 공식을 혼동하거나, 중심화된 모형과 중심화되지 않은 일반 모형의 절편 분산 공식을 그대로 바꿔 쓰는 것이다. $\frac{(\bar{x})^2}{\sum(x_i - \bar{x})^2}$ 항이 붙는 공식은 설명변수가 x_i 그대로 들어간 모형에서만 성립하며, 이 문제처럼 $(x_i - \bar{x})$로 중심화된 모형에서는 그 항이 사라지고 $Var(\hat{\beta}_0) = \sigma^2/n$이 된다는 점을 놓치면 안 된다.

다시 나오면 이렇게 판단

회귀모형에서 설명변수가 x_i로 그대로 있는지 $(x_i - \bar{x})$로 중심화되어 있는지를 먼저 확인한다. 중심화된 모형이면 $\hat{\beta}_0 = \bar{Y}$이므로 $Var(\hat{\beta}_0) = \sigma^2/n$이고, 중심화되지 않은 일반 모형이면 $Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)$이 된다는 두 공식을 짝지어 기억해 둔다. 기울기 추정량의 분산 $Var(\hat{\beta}_1) = \sigma^2/\sum(x_i - \bar{x})^2$은 두 경우 모두 동일하다.

암기 포인트

중심화 모형($x_i - \bar{x}$ 사용): $\hat{\beta}_0 = \bar{Y}$, $Var(\hat{\beta}_0) = \sigma^2/n$. 일반 모형($x_i$ 그대로 사용): $Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)$. 기울기 분산 $Var(\hat{\beta}_1) = \sigma^2/\sum(x_i - \bar{x})^2$은 공통.

NoteBase

문 22. 공정 1, 2, 3에서 생산된 제품의 강도에 차이가 있는지를 알아보기 위해 세 공정에서 생산된 제품을 임의로 선택하고 강도를 측정하여 요약한 결과가 다음과 같을 때, 이를 이용한 분산분석의 설명으로 옳지 않은 것은? (단, n_i는 공정 i에서 선택된 제품 수, y_{ij}는 공정 i에서 j번째 관측 강도, $\bar{y}_{i\cdot}$는 공정 i에서의 평균 강도, $\bar{y}$는 강도의 전체 평균을 나타낸다)

공정(i)	1	2	3
n_i	4	5	6
$\bar{y}_{i\cdot}$	117	120	122
$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$	60	80	100
$\bar{y}$	120		
$\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$	300		

① 총제곱합의 자유도는 14이다.

② 오차제곱합은 240이다.

③ 공정에 대한 평균제곱은 20이다.

④ 귀무가설하에서 검정통계량의 분포는 자유도가 (2, 12)인 F-분포이다.

정답 3번

출제 포인트

일원배치 분산분석표를 직접 채워 완성할 수 있는지를 묻는 문항이다. 표에 주어진 각 공정의 표본크기 n_i, 표본평균 $\bar{y}_{i\cdot}$, 공정 내 편차제곱합 $\sum(y_{ij}-\bar{y}_{i\cdot})^2$, 전체평균 $\bar{y}$, 총제곱합 $\sum\sum(y_{ij}-\bar{y})^2$을 이용해 처리제곱합(공정간제곱합), 자유도, 평균제곱, 검정통계량의 분포를 각각 계산해 옳고 그름을 판단해야 한다.

조건 정리

공정 수(처리 수) $k=3$, 각 공정의 표본크기는 $n_1=4,\ n_2=5,\ n_3=6$이므로 전체 표본크기 $N=n_1+n_2+n_3=15$이다. 공정별 평균은 $\bar{y}_{1\cdot}=117,\ \bar{y}_{2\cdot}=120,\ \bar{y}_{3\cdot}=122$이고 전체 평균은 $\bar{y}=120$이다. 공정 내 제곱합은 각각 60, 80, 100이고 총제곱합(SST)은 300이다.

NoteBase

사용할 공식

총제곱합은 처리제곱합과 오차제곱합으로 분해된다: $SST = SSTr + SSE$. 오차제곱합은 각 공정 내 편차제곱합의 합이다: $SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{i.})^2$. 처리제곱합은 $SSTr = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y})^2$로도 구할 수 있다. 자유도는 총제곱합 $N-1$, 처리제곱합 $k-1$, 오차제곱합 $N-k$이며, 각 평균제곱은 제곱합을 해당 자유도로 나눈 값이다. 귀무가설(모든 공정의 평균이 같다)하에서 검정통계량은 $F = \frac{MSTr}{MSE}$이고 자유도가 $(k-1,\ N-k)$인 F-분포를 따른다.

계산 과정

총자유도는 $N-1 = 15-1 = 14$이다. 오차제곱합은 표에 제시된 세 값을 더하면 되므로 $SSE = 60+80+100 = 240$이다. 처리제곱합은 $SST - SSE = 300 - 240 = 60$이며, 직접 계산해도 $SSTr = 4(117-120)^2 + 5(120-120)^2 + 6(122-120)^2 = 4\times 9 + 5\times 0 + 6\times 4 = 36+0+24 = 60$으로 같다. 처리(공정)의 자유도는 $k-1 = 3-1 = 2$이므로 공정에 대한 평균제곱은 $MSTr = \frac{SSTr}{k-1} = \frac{60}{2} = 30$이다. 오차자유도는 $N-k = 15-3 = 12$이므로 귀무가설하 검정통계량은 자유도 $(2, 12)$인 F-분포를 따른다.

선지 판단

① 총제곱합의 자유도는 14이다 → O. $N-1 = 15-1 = 14$로 계산 결과와 일치한다.
② 오차제곱합은 240이다 → O. 세 공정의 공정 내 편차제곱합을 더한 $60+80+100 = 240$과 일치한다.
③ 공정에 대한 평균제곱은 20이다 → X. 실제로는 $MSTr = SSTr/(k-1) = 60/2 = 30$이므로 20이 아니라 30이다. 틀린 표현 → 올바른 표현: "평균제곱은 20이다" → "평균제곱은 30이다".
④ 귀무가설하에서 검정통계량의 분포는 자유도가 (2,12)인 F-분포이다 → O. 처리자유도 $k-1 = 2$, 오차자유도 $N-k = 12$이므로 옳다.

오답 포인트

SSTr를 구할 때 처리자유도 $k-1 = 2$로 나누어야 하는데, 총자유도 $N-1 = 14$나 표본크기와 혼동해 다른 값으로 나누면 20 같은 엉뚱한 숫자가 나올 수 있다. 또한 총제곱합에서 오차제곱합을 빼서 처리제곱합을 구하는 분해식 $SST = SSTr + SSE$를 기억하지 못하면 60이라는 처리제곱합 자체를 얻지 못한다.

다시 나오면 이렇게 판단

표 형태의 일원배치 자료가 나오면 먼저 $N = \sum n_i$, k(처리 수)를 확인하고, SSE=공정 내 제곱합의 합, $SSTr = SST - SSE$(또는 $\sum n_i(\bar{y}_{i.} - \bar{y})^2$)로 각각 구한다. 그 다음 자유도는 총 $N-1$, 처리 $k-1$, 오차 $N-k$로 나누고, 평균제곱은 반드시 제곱합을 해당 자유도로 나눈 값인지 검산한다.

암기 포인트

일원배치 분산분석: $SST = SSTr + SSE$, 자유도 $N-1 = (k-1) + (N-k)$, $MSTr = SSTr/(k-1)$, $MSE = SSE/(N-k)$, $F = MSTr/MSE \sim F(k-1,\ N-k)$.

NoteBase

문 23. 다음은 납축전지의 효율이 온도와 습도 요인에 따라 영향을 받는지를 알아보기 위한 실험의 분석 결과를 정리한 분산분석표의 일부이다. 실험은 3가지 온도와 4가지 습도의 모든 수준 조합에서 한 번씩 이루어졌다. 이에 대한 설명으로 옳은 것은?

요인	제곱합	자유도	$F-$값	$p-$값
온도	42			0.027
습도	18			0.216
오차	18			

① 오차항의 자유도는 8이다.

② 습도에 대한 F-값은 4이다.

③ 유의수준 5%에서 납축전지의 효율이 온도의 수준에 따라 모두 같다고 할 수 있다.

④ F가 자유도 (2,6)인 F-분포를 따르는 확률변수일 때, 온도에 대한 p-값은 $P(F > 7)$이다.

정답 4번

출제 포인트

반복이 없는 이원배치 분산분석표를 완성할 수 있는지를 묻는 문항이다. 온도(3수준)와 습도(4수준)의 모든 조합에서 각각 한 번씩만 실험했으므로 반복이 없는 이원배치 설계이며, 이 경우 오차의 자유도는 두 요인 자유도의 곱으로 결정된다. 표에 비어 있는 자유도·F-값을 직접 채워야 각 선지의 참거짓을 판단할 수 있다.

조건 정리

온도 수준수 $a = 3$, 습도 수준수 $b = 4$, 반복 없이 모든 조합에서 1회씩 실험. 온도 제곱합 $SS_A = 42$, 습도 제곱합 $SS_B = 18$, 오차 제곱합 $SS_E = 18$. 온도 p-값 0.027, 습도 p-값 0.216이 주어져 있다.

사용할 공식

반복 없는 이원배치에서 자유도는 온도 $df_A = a - 1$, 습도 $df_B = b - 1$, 오차 $df_E = (a-1)(b-1)$, 전체 $df_T = ab - 1$이다. 각 요인의 평균제곱은 $MS = \frac{SS}{df}$이고, F=MS_요인MS_E로 계산하며, 이 F는 자유도 (df_요인, df_E)인 F-분포를 따른다.

NoteBase

계산 과정

자유도: $df_A = 3-1 = 2$, $df_B = 4-1 = 3$, $df_E = (3-1)(4-1) = 6$이고 합은
$2+3+6 = 11 = ab-1 = 3\times 4-1$로 전체 자유도와 일치한다.
평균제곱: $MS_A = \frac{42}{2} = 21$, $MS_B = \frac{18}{3} = 6$, $MS_E = \frac{18}{6} = 3$.
F-값: 온도 $F_A = \frac{21}{3} = 7$, 습도 $F_B = \frac{6}{3} = 2$.
따라서 온도에 대한 검정통계량은 자유도 (2, 6)인 F-분포를 따르고 관측된 값은 7이므로, 이 F-값에 대응하는 p-값은 $P(F > 7)$이며 이는 표에 제시된 0.027과 대응한다.

선지 판단

① 오차항의 자유도는 8이다 → X. 오차 자유도는 $(3-1)(4-1) = 6$이지 8이 아니다.
② 습도에 대한 F-값은 4이다 → X. $MS_B = 6$, $MS_E = 3$이므로 $F_B = 6/3 = 2$이다.
③ 유의수준 5%에서 납축전지의 효율이 온도의 수준에 따라 모두 같다고 할 수 있다 → X. 온도의 p-값 0.027은 유의수준 0.05보다 작으므로 귀무가설(온도 수준 간 효율이 모두 같다)을 기각한다. 즉 온도 수준에 따라 효율이 다르다고 판단해야 한다.
④ F가 자유도 (2,6)인 F-분포를 따르는 확률변수일 때, 온도에 대한 p-값은 P(F>7)이다 → O. 계산 과정에서 구한 대로 온도의 검정통계량은 자유도 (2, 6)인 F-분포를 따르고 관측값이 7이므로 p-값은 $P(F > 7)$이 맞다.

오답 포인트

반복이 없는 이원배치에서 오차 자유도를 전체 자유도에서 두 요인 자유도만 빼는 것으로 착각하면 8(=11-3)처럼 잘못 계산하기 쉬운데, 두 요인의 자유도를 더 빼야 한다는 점(11-2-3=6)을 놓치지 않아야 한다. 또한 F-값은 반드시 요인의 평균제곱을 오차의 평균제곱으로 나눈 값이지 제곱합끼리 비교한 값이 아니라는 점, p-값이 작을수록 귀무가설을 기각(수준 간 차이가 있다고 판단)한다는 방향을 혼동하지 않아야 한다.

다시 나오면 이렇게 판단

반복 없는 이원배치 분산분석표 문제는 항상 자유도부터 채운다. 요인 자유도는 수준수-1, 오차 자유도는 두 요인 자유도의 곱, 전체 자유도는 이 셋의 합이자 전체관측수-1과 같은지로 검산한다. 그다음 각 요인의 평균제곱과 F-값을 구하고, p-값과 유의수준을 비교해 귀무가설 기각/채택 방향을 판단한다.

암기 포인트

반복 없는 이원배치: $df_A = a-1$, $df_B = b-1$, $df_E = (a-1)(b-1)$, $df_T = ab-1$. F=MS_요인MS_E이며 p-값이 유의수준보다 작으면 그 요인의 효과가 있다고 판단(수준 간 차이가 있다)한다.

NoteBase

문 24. 모수 θ에 따른 이산형 확률변수 X의 분포 $f(x;\theta)$는 다음과 같다. 위 분포로부터 크기 2인 확률표본 X_1과 X_2가 각각 1과 3으로 관측되었을 때, θ에 대한 최대가능도추정치(maximum likelihood estimate)는?

$f(x;\theta)$ \ θ	$\theta=0$	$\theta=2$	$\theta=4$	$\theta=6$
$f(1;\theta)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{5}$
$f(2;\theta)$	$\frac{1}{3}$	0	$\frac{1}{4}$	$\frac{2}{5}$
$f(3;\theta)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{2}{5}$

① 0

② 2

③ 4

④ 6

정답 2번

출제 포인트

이산형 확률변수의 결합분포표가 주어졌을 때, 관측값에 대응하는 우도함수 $L(\theta)=f(1;\theta)\times f(3;\theta)$를 θ의 각 후보값(0, 2, 4, 6)에 대해 직접 계산하여 비교하는 문제이다. 최대가능도추정치(MLE)는 우도함수 값을 가장 크게 만드는 θ이다.

조건 정리

표본 $X_1=1, X_2=3$이 독립적으로 관측되었으므로 결합우도는 $L(\theta)=f(1;\theta)\cdot f(3;\theta)$이다. 표에서 θ별 $f(1;\theta)$와 $f(3;\theta)$ 값을 읽으면 다음과 같다.
$\theta=0$: $f(1;0)=\frac{1}{3}$, $f(3;0)=\frac{1}{3}$
$\theta=2$: $f(1;2)=\frac{1}{2}$, $f(3;2)=\frac{1}{2}$
$\theta=4$: $f(1;4)=0$, $f(3;4)=\frac{3}{4}$
$\theta=6$: $f(1;6)=\frac{1}{5}$, $f(3;6)=\frac{2}{5}$

NoteBase

사용할 공식

독립인 두 관측값의 결합우도는 각 관측값의 확률(질량함수 값)의 곱이다.
$L(\theta) = f(1;\theta) \times f(3;\theta)$
MLE는 이 값을 최대로 만드는 θ이다.

계산 과정

각 θ에 대해 우도를 계산한다.
$\theta = 0$: $L(0) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9} \approx 0.111$
$\theta = 2$: $L(2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0.25$
$\theta = 4$: $L(4) = 0 \times \frac{3}{4} = 0$
$\theta = 6$: $L(6) = \frac{1}{5} \times \frac{2}{5} = \frac{2}{25} = 0.08$
네 값을 비교하면 $L(2) = 0.25$가 가장 크다. 따라서 우도를 최대화하는 θ는 2이다.

선지 판단

① 0: $L(0) = 1/9 \approx 0.111$로 최댓값이 아니다. X
② 2: $L(2) = 1/4 = 0.25$로 네 값 중 가장 크다. O
③ 4: $f(1;4) = 0$이므로 $L(4) = 0$이 되어 관측값 $X_1 = 1$이 나올 확률 자체가 0이다. X
④ 6: $L(6) = 2/25 = 0.08$로 $L(2)$보다 작다. X

오답 포인트

이 문제에서 흔히 하는 실수는 $f(1;\theta)$와 $f(3;\theta)$ 중 하나만 보고 큰 값을 고르는 것이다. 예를 들어 $f(3;4) = 3/4$가 표에서 가장 큰 값이라고 해서 ③번을 고르면 틀린다. $\theta = 4$에서는 $f(1;4) = 0$이므로 결합우도 전체가 0이 되어 버린다. 반드시 두 관측값의 확률을 모두 곱한 결합우도를 비교해야 한다.

다시 나오면 이렇게 판단

이산형 결합분포표에서 여러 표본이 동시에 관측된 경우의 MLE를 묻는 문제는, 먼저 각 관측값에 해당하는 행을 표에서 확인한 뒤 후보 θ마다 그 값들을 모두 곱해 우도를 구하고, 그 중 최댓값을 갖는 θ를 고른다. 표의 한 행이나 한 칸만 보고 판단하지 않는다.

암기 포인트

여러 개의 독립인 표본이 있을 때 우도함수는 각 표본의 확률(질량함수 값)을 모두 곱한 것이며, 어느 하나라도 확률이 0이면 그 θ에서의 결합우도는 0이 된다는 점을 기억해 둔다.

NoteBase

문 25. 두 연속형 확률변수 X와 Y의 결합확률밀도함수가 다음과 같을 때, $P(X+Y \le 1)$의 값은?

f(x,y)=cases(1)/(2), & 0 ≤ x ≤ 1, 0 ≤ y ≤ 2 \ 0, & 그 외cases

① $\frac{1}{2}$
② $\frac{1}{3}$
③ $\frac{1}{4}$
④ $\frac{1}{6}$

정답 3번

출제 포인트

이 문제는 두 변수의 결합확률밀도함수가 직사각형 영역에서 균등(uniform)하게 주어질 때, 사건 X+Y≤1의 확률을 이중적분(또는 넓이 비율)으로 구하는 능력을 확인한다. 핵심은 지지영역과 조건 x+y≤1이 만나는 부분의 넓이를 정확히 찾아내는 것이다.

조건 정리

$f(x,y)=\frac{1}{2}, 0 \le x \le 1,\ 0 \le y \le 2$이고 그 외에는 0이다. 지지영역은 가로 1, 세로 2인 직사각형이며 넓이는 2이다. 밀도가 상수 $\frac{1}{2}$이므로 이는 균등분포이고, 전체 확률은 $\frac{1}{2} \times 2 = 1$로 정상화되어 있다. 구해야 할 사건은 $X+Y \le 1$이며, 이는 직선 $x+y=1$ 아래쪽 영역과 지지영역의 교집합이다.

사용할 공식

균등분포에서는 확률이 밀도 상수와 해당 영역 넓이의 곱으로 구해진다. P(X+Y≤ 1)=_A f(x,y) dx dy = (1)/(2)×(영역 A의 넓이) 여기서 영역 A는 지지영역 $\{0 \le x \le 1,\ 0 \le y \le 2\}$와 $\{x+y \le 1,\ x \ge 0,\ y \ge 0\}$의 교집합이다.

계산 과정

직선 $x+y=1$은 $x=0$일 때 $y=1$, $y=0$일 때 $x=1$을 지난다. x의 범위가 [0, 1]로 제한되어 있으므로 각 x에 대해 y의 상한은 $1-x$이며, $0 \le x \le 1$에서 $1-x$는 항상 [0, 1] 사이의 값이라 지지영역의 상한인 $y \le 2$보다 작다. 따라서 조건 $x+y \le 1$은 지지영역 안에서 그대로 유효하며, 교집합 영역 A는 꼭짓점이 (0, 0), (1, 0), (0, 1)인 직각삼각형이다. 이 삼각형의 넓이는 (1)/(2)×(밑변)×(높이)=(1)/(2)× 1× 1=(1)/(2) 이므로 $P(X+Y \le 1)=\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$이다.

NoteBase

선지 판단

① 1/2 → X: 전체 지지영역 넓이(2)에 밀도를 곱한 값(=1)의 절반과 혼동하거나, 삼각형 넓이 비율을 잘못 잡은 경우 나오는 값이다. 옳은 계산과 다르다.
② 1/3 → X: 삼각형 대신 다른 비율(예: 밑변과 높이를 다르게 잡은 경우)로 계산했을 때 나올 수 있는 값이며, 이 문제의 영역 넓이와 맞지 않는다.
③ 1/4 → O: 밀도 $\frac{1}{2}$과 삼각형 넓이 $\frac{1}{2}$을 곱한 $\frac{1}{4}$이 정확한 값이다.
④ 1/6 → X: 삼각형 넓이를 $\frac{1}{3}$로 잘못 구하거나 밑변·높이를 잘못 곱했을 때 나오는 값으로, 실제 교집합 영역과 맞지 않는다.

오답 포인트

이 문제에서 가장 흔한 실수는 지지영역이 $y \le 2$까지라는 것을 보고 삼각형이 아니라 사각형이나 다른 도형으로 잘못 판단하는 것이다. 하지만 x의 범위가 $[0, 1]$로 제한되어 있어 직선 $x + y = 1$이 만드는 삼각형은 항상 $y \le 1$ 범위 안에 있으므로 $y \le 2$ 조건은 실제로 이 삼각형을 자르지 않는다. 이 점을 놓치면 넓이를 다르게 계산해 오답을 고르게 된다.

다시 나오면 이렇게 판단

결합확률밀도함수가 직사각형 위에서 상수로 주어지는 균등분포 문제는, 먼저 지지영역의 전체 넓이를 확인하고 밀도 상수가 넓이의 역수인지(정상화 조건) 검산한다. 그 다음 구하려는 사건의 부등식이 그리는 직선이나 곡선이 지지영역 내부에서 어떤 도형(삼각형, 사다리꼴 등)을 만드는지 그려보고, 그 도형의 넓이에 밀도 상수를 곱해 확률을 구한다.

암기 포인트

균등분포에서 확률 = 밀도 상수 × 해당 영역의 넓이라는 원리를 기억해 둔다. 또한 부등식으로 주어진 영역이 지지영역 전체를 넘어서는지, 지지영역 안에 완전히 포함되는지를 먼저 확인하는 습관을 들여 둔다.

NoteBase